Mind Out of Matter

Topics in the Physical Foundations of Consciousness and Cognition

Gregory R. Mulhauser

Ph.D.
The University of Edinburgh
1994

Dedicated to the memory of my grandfather,
Raymond B. Mulhauser,
who died half a century after fighting to liberate North Africa and
Europe and exactly two months before this dissertation
was completed on free European soil.

Table of Contents

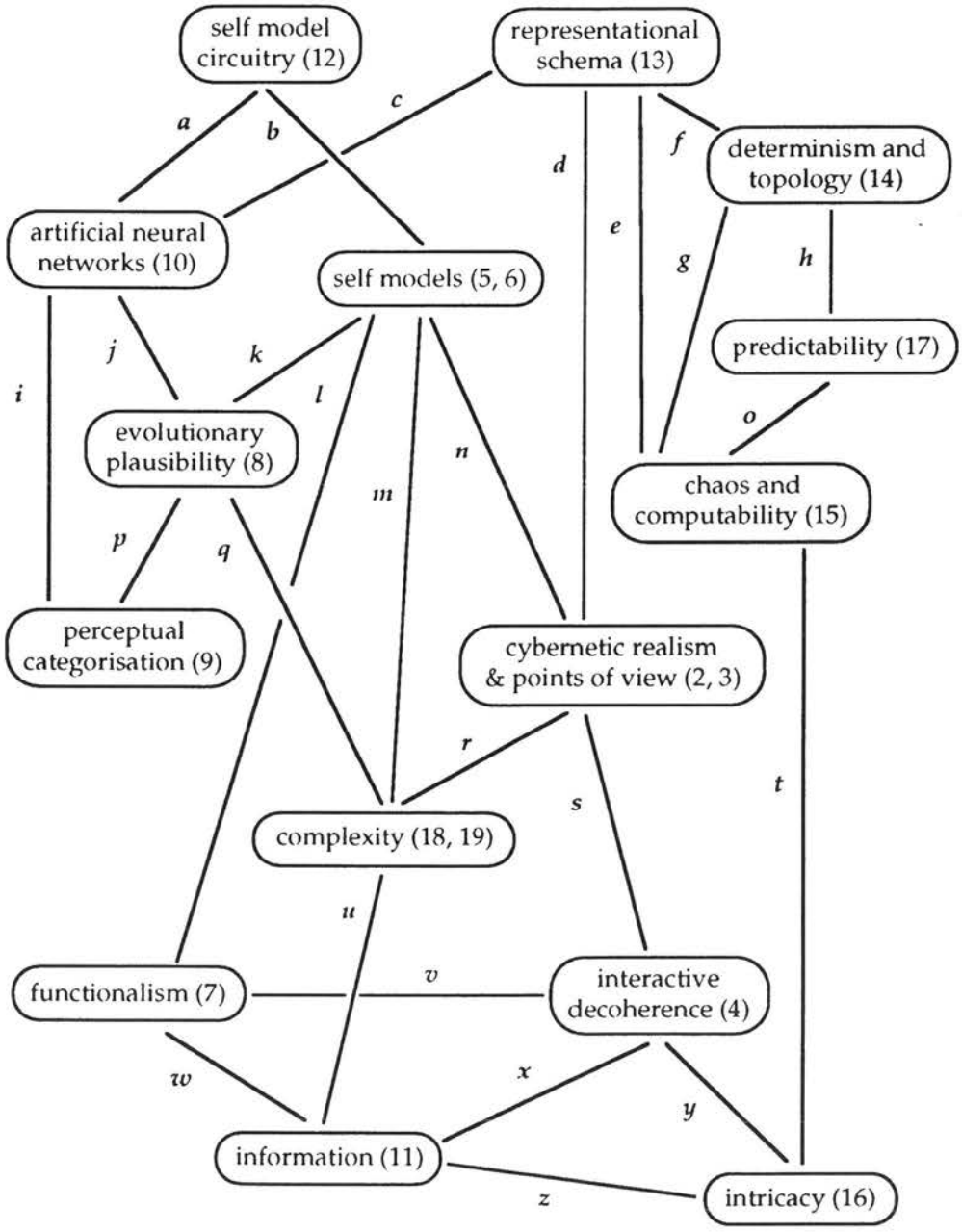
Abstract

This dissertation begins with an exploration of a brand of dual aspect monism and some problems deriving from the distinction between a first person and third person point of view. I continue with an outline of one way in which the conscious experience of the subject might arise from organisational properties of a material substrate. With this picture to hand, I first examine theoretical features at the level of brain organisation which may be required to support conscious experience and then discuss what bearing some actual attributes of biological brains might have on such experience. I conclude the first half of the dissertation with comments on information processing and with artificial neural networks meant to display simple varieties of the organisational features initially described abstractly.

While the first half begins with a view of conscious experience and infers downwards in the organisational hierarchy to explore neural features suggested by the view, attention in the second half shifts towards analysing low level dynamical features of material substrates and inferring upwards to possible effects on experience. There is particular emphasis on clarifying the rôle of chaotic dynamics, and I discuss relationships between levels of description of a cognitive system and comment on issues of complexity, computability, and predictability before returning to the topic of representation which earlier played a central part in isolating features of brain organisation which may underlie conscious experience.

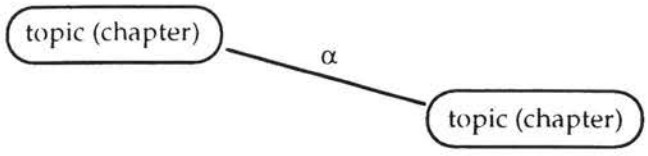
Some themes run throughout the dissertation, including an emphasis on understanding experience from both the first person and the third person points of view and on analysing the latter at different levels of description. Other themes include a sustained effort to integrate the picture offered here with existing empirical data and to situate current problems in the philosophy of mind within the new framework, as well as an appeal to tools from mathematics, computer science, and cognitive science to complement the more standard philosophical repertoire.

Themes and Connections



(key on two pages following)

Themes and Connections Key



α— nature of connection

a— the artificial neural network framework of Chapter 10 is the basis of the self model circuitry we explore in Chapter 12

b— some example properties of self models from Chapters 5 and 6 are implemented in Chapter 12

c— the schema of Chapter 13 offers a way of tracing self model circuitry as a dynamical system and of relating the dynamics at this implementational level to dynamics at the level of psychological transitions

d— the schema of Chapter 13 offers a more formal framework for the levels of description and points of view which were crucial to Chapters 2 and 3

e— the possibility of chaotic dynamics was one motivation for the schema of Chapter 13, and Chapter 15 offers an example of an important λ level influence on the ψ level

f— the schema of Chapter 13 is a good framework for understanding the argument of Chapter 14 and the defence against Smith

g— Chapters 13 and 15 each offer different arguments for problems with ψ level computability or determinism

h— Chapter 17 includes an analysis of predictability problems which bear on ψ indeterminism as discussed in Chapter 14

i— the artificial neural network examples of Chapter 10 feature some characteristics of Edelman's perceptual categorisation work

j— our concern with evolutionary plausibility in Chapter 8 features again in an example network of Chapter 10

k— in Chapter 8, some of the features of self models described in Chapters 5 and 6 are justified in an evolutionary context

l— Chapter 7 explores some relationships between functionalism and the self model properties of Chapters 5 and 6

m— the notion of representation used in Chapters 5 and 6 is refined in Chapter 19

n— the self model view is one approach to describing the third person aspect of conscious awareness in the cybernetic realism framework of Chapters 2 and 3

o— Chapter 17 offers a real example of the behaviour predicted in Chapter 15

p— in keeping with the evolutionary emphasis of Chapter 8, Chapter 9 describes Edelman's biologically plausible account of perceptual categorisation with "evolution" on a somatic time scale

q— Chapter 19 illustrates possible biologically plausible rôles for chaotic signals which are consistent with evolutionary constraints

r— Chapter 19 reveals the importance of points of view, so important to Chapter 2 and 3, to evaluating complexity

s— Chapter 4 secures the approach developed in Chapters 2 and 3 against criticisms from quantum mechanics

t— given the importance of the conclusions of Chapter 15, Chapter 16 begins a three chapter defence of the relevance of chaos (for simplicity, only Chapter 16 is shown connected to Chapter 15)

u— points about complexity and noise from Chapters 18 and 19 are closely related to the definition of information which we explored in Chapter 11

v— if Chapter 4 is wrong, the functionalism of Chapter 7 as well as the advantages which the self model view has over it are in peril

w— the definition of information of Chapter 11 accommodates distributed and functionalist notions of processing

x— Chapters 4 and 11 both appeal to a definition of information strictly dependent on states of physical systems

y— the interactive decoherence of Chapter 4 bears on the limits to intricacy in physical systems

z— the rôle of measurable information (Chapter 10) about a physical system is crucial to understanding our points about infinite intricacy in models and in the world

Foreword and Acknowledgements

While I take full responsibility for what has finally been written, producing this dissertation has benefitted from the support of a long list of organisations and individuals. Foremost among these is the British Marshall Aid Commemoration Commission, administered under the auspices of the Association of Commonwealth Universities, without whose three years of financial support for my living expenses and research at the University of Edinburgh this dissertation would likely never have been written. For grants and bursaries assisting with travel and subsistence expenses, I also thank the following individuals and organisations:

- Analysis Trust
- British Society for the Philosophy of Science
- Oxford Centre for the Environment, Ethics, and Society
- Commission of the European Communities Human Capital and Mobility Programme
- Society for the Study of Artificial Intelligence and Simulation of Behaviour
- University of Edinburgh Cognitive Science Centre
- Ray Mulhauser (deceased) and Florence Mulhauser
- University of Edinburgh Department of Philosophy
- University of Edinburgh Human Communication Research Centre

To be sure, funding does not begin to tell the whole story of completing such a project, and the more individual part of that story begins with my first primary supervisor, Stig Rasmussen, and his successor, Alexander Bird, in addition to the Head of Department Willie Charlton and the secretary Seona Macintosh. I appreciate the time of professional correspondents who have kept me supplied with especially important new papers and/or challenged my thinking on substantive points; these include Sue Blackmore, Tony Browne, James Garson, Jesse Hobbs, Terry Horgan, Brian Josephson, Thomas Metzinger, Adam Morton, David Rumelhart, and especially Peter Smith, who has managed to write more comments about parts of this material than any other single person. I am grateful also to Pat Churchland and Christof Koch for some brief but

much needed early encouragement on hideous drafts of a paper which grew into the chaos and computability material of Chapter 15. I am indebted both to the staff of the International Philosophical Preprint Exchange at Japan's Chiba University and to the numerous readers from across the world who have downloaded portions of this text from that service or from its dozens of mirror sites, as well as to anonymous referees at the journal *Minds and Machines* and referees and/or conference and seminar audiences in Scotland, England, Belgium, Russia, and Poland.

Finally, and most dear to my heart, are my parents, Gerald and Dawn Mulhauser, and my friends, without the support of whom I should not have lasted through to the end of this project. The friends and acquaintances who have offered their support, love, and inspiration (in various proportions!) are too numerous to mention in full, and I hope none will be offended if I have forgotten to include them here. Among these people I am especially grateful to Violet Anderson, my kobudo and karate sempai Eva Bayerlein and Andrew Horne, Jeff Carr, Olga Illicholva, Till-Christof Jelitto, Pascal Lynch, Mags MacLean, Kristina Mathesen, Jim McCormick, Julia Shaffner, my kobudo sensei John Tosney, and Rachel Weaver. Thanks to them also for putting up with me and to my girlfriend Hélène Marie for her love even in the face of my anti-social finish-the-dissertation psychological spasming.

Gregory R. Mulhauser
Edinburgh, Scotland
13 September 1994

Please note:

At several points in this text, I have cited evidence obtained from neurophysiological studies of animals. Often these studies have been performed on primates or other mammals, and often they have involved in vivo *lesioning, ablation, or downright mutilation of the animals concerned. Although the data obtained through some of these studies have fuelled the development of my own theories, my referencing the studies does not indicate that I support invasive* in vivo *procedures. In general, I do not sanction such procedures, and I do not intend my work to spur the collection of any other data in such fashion. Under the influence of Singer, Regan, and others, I deplore this exploitation, even if the relevant knowledge could not have been had any other way. Knowledge that demands such abuse is, to my mind, not worth having.*

1

Introduction

To ask or search I blame thee not; for Heaven
is as the Book of God before thee set,
Wherein to read His wondrous works, and learn
His seasons, hours, or days, or months, or years.

Paradise Lost, Book 8

1.1 Science and the Loss of Mystery

One of the first philosophy texts I encountered as a first year undergraduate was titled simply *Metaphysics*. There, Richard Taylor told the story of a poor fellow called Osmo, who perished trying to run away from a future foretold for him in a special book which described every event in his entire life, from his birth to his death. Like the protagonist in a Greek tragedy, Osmo could not escape his Fate: what was written in the book *always* happened, and try as he might, his every struggle to forge a different future served only to convince him yet more strongly of the book's infallibility.

Taylor's point in telling the story, apart perhaps from terrorising nascent young philosophers, was that if it were logically possible that such a complete and true biographical book had *already* been written about us—before we had got even a third of the way through our lives—then we really oughtn't fret over any of those things in the book which it was already true were going to happen to us anyway. That is, since all the descriptions of *future* events in our lives are true right *now*, and since our complete stories could already have been written in a book, we all should be logical fatalists and stop worrying about a future which it is true now will come to pass. Of course—despite Taylor's protests to the contrary—

the reasoning is modally fallacious. But then, as now, I can almost see what made Taylor suggest the line of thought, despite his comprehensive understanding of the logic of the modal fallacy. Most philosophers reject Taylor's reasoning, yet somehow there is still something faintly disturbing about the idea that somewhere there could be a book accurately foretelling my every experience for the rest of my life. I suspect that peculiar uncomfortableness might be akin to the distaste some people have for the idea that human cognition and conscious sensation may one day be explained by science.

Most of us cling more or less tenaciously to some idea of privacy of our own innermost mental experiences, and the possibility of there being a book which could parade about for everyone to see the every outcome of every deliberation, the every reaction to every event in our lives—before any of it has even happened—threatens that sense of a peaceful inner retreat. Likewise, perhaps many abhor the thought of some brain researchers invading this mind space, our last bastion of privacy, pushing back the last frontiers of scientific inquiry and explaining how our experience arises from the brain, the most complex object in the humanly known cosmos. For centuries, humans have held to an ontology of strict dualism which separated mind things from physical things, and while the physical world yielded progressively more of its secrets to the piercing eyes of empirical science, the world of mind and spirit lay safely beyond its reach. Now our modern world is firmly in the grip of material monism, and many I believe fear that it will banish forever whatever comforts humans once took from their dualist paradigm, exposing life and human experience as mere vibrations of particles in the quantum vacuum—and as valueless. Scientific explanation, the unacknowledged suspicion seems to say, goes hand in hand with the loss of mystery, the loss of wonder, and the loss of some kind of value of whatever is being explained.

Yet a mind by any other name—or scientific description—is still as wonderful a thing! A person still wonders, still hurts, still laughs, still falls in love, regardless of whether those things all are achieved solely by a brain made of particles vibrating in the quantum vacuum. Knowing what chemicals make up the surfaces of the paintings in the Rembrandt room at St. Petersburg's Hermitage renders them no less dark and brooding. Indeed, perhaps a complete material explanation of the reflection of light from a Rembrandt onto my retina and through the various pathways in

my brain, culminating in my experience of the painting would offer an even *more* amazing angle on the whole affair. If *I* really am just a collection of vibrating particles, isn't it all the more incredible that I can have my experiences at all, that I do *see* anything in the mix of colours, that I gain some insight, imagined or otherwise, into the mind of an artist who painted centuries before I was born? Far from the encounter robbing us of the mystery and sense of privacy of mental experience, when scientific explanation meets conscious experience, the mystery of the latter is not eroded but is instead rendered all the more impressive. I believe a purely material science of the mind/brain can do nothing but enrich that experience and offer us the beginning of an integrated understanding of ourselves and our place in the context of a material cosmos. It is to furthering this cause of a purely material science of the mind/brain that I am in this dissertation committed.

1.2 General Themes and Methodology

Our first aim is to explore a brand of dual aspect monism and some problems deriving from the distinction between a first person and third person point of view. We continue by outlining one way in which the conscious experience of the subject might arise from organisational properties of a material substrate. With this picture to hand, we first examine theoretical features at the level of brain organisation which may be required to support conscious experience and then discuss what bearing some actual attributes of biological brains might have on such experience. We conclude the first half of the dissertation with comments on information processing and with artificial neural networks meant to display simple varieties of the organisational features initially described abstractly.

While the first half begins with a view of conscious experience and infers downwards in the organisational hierarchy to explore neural features suggested by the view, attention in the second half shifts towards analysing low level dynamical features of material substrates and inferring upwards to possible effects on experience. There is particular emphasis on clarifying the rôle of chaotic dynamics, and I discuss relationships between levels of description of a cognitive system and comment on issues of complexity, computability, and predictability before returning to the topic

of representation which earlier played a central part in isolating features of brain organisation which may underlie conscious experience.

Some themes run throughout the dissertation, including an emphasis on understanding experience from both the first person and the third person points of view and on analysing the latter at different levels of description. Other themes include a sustained effort to integrate the picture offered here with existing empirical data and to situate current problems in the philosophy of mind within the new framework, as well as an appeal to tools from mathematics, computer science, and cognitive science to complement the more standard philosophical repertoire.

The overall structure of this dissertation mirrors the suggestion I make to students about good essay writing: tell them what you're going to tell them, tell them, and then tell them what you've told them. The first and the last bits correspond to Chapters 1 and 20, respectively; it's the other chapters which take some time to get through, and it's the rest for which I'll offer a sort of road map in the next section. Before moving on to a quick summary of the individual chapters, however, it is first helpful to make a brief note about methodology. In what follows, our strategy is to sketch positions on a number of foundational issues and then try to exploit those positions to understand something of the relationship between the wetware of the brain and the experience of a conscious subject. We don't engage in the Quixotic task of attempting to prove that each of these positions is the one and only correct approach to the matter at hand. Instead, we explore the *prima facie* plausibility of an approach, treat it essentially as a *fait accompli*, and then get on with exploring what philosophical problems might be made more or less palatable by the view on offer. (This is not to say, of course, that we neglect altogether the task of ensuring the frameworks within which our discussions and explorations take place are coherent and internally consistent.) We may then look back at the positions we have taken on foundational issues with a better understanding of whether they are useful positions worthy of further attention.

Our methodology is akin to that of the physicist who proceeds as follows. Suppose the weak nuclear force were mediated by the exchange of things I'll call w-particles... What would that supposition allow me to explain? What would be the other consequences of positing w-particles? Is it reasonable to suppose w-particles exist, given what I already know and

given the other hypotheses to which I'm particularly attached? With what existing data and hypotheses would it be especially important to check that the supposition that w-particles exist is consistent? As more of these questions are answered, the physicist learns whether it helps her understanding of the world to suppose w-particles exist, and she learns whether positing them requires modifying any other parts of her model of the world. What she does *not* do is begin by attempting to prove from existing data that w-particles *must* exist. Like the physicist, by the time we have finished mimicking this process several times, we will hopefully have made some progress toward understanding whether our initial positions on foundational issues are useful.

1.3 Chapters Summary

This process begins with our adopting a brand of dual aspect monism we dub "cybernetic realism". The central tenet of this line is that within the context of material monism, we may still hold that there are matters of fact about what it is like to *be* a given material thing which may not be expressible purely in terms of the objective physical properties of that thing. We introduce this view first with a brief reprinted paper commenting on Nagel's famous question, "what is it like to be a bat?", and then elucidate it in a followup chapter where two central themes underlying the whole of this dissertation first emerge most clearly. The first of these central themes is our concern for examining and modelling cognitive systems at different levels of description and noticing what sorts of properties may be observed at the highest, intermediate, and lowest levels. The second theme is our emphasis on understanding cognitive systems from both the third person point of view and the point of view of the first person conscious subject.

With this brand of dual aspect monism to hand, we turn our attention next to answering a side concern which might pose a problem for materialism itself and especially for any theory of mind which is couched strictly in the terms of materialism. This chapter is based on a forthcoming *Minds and Machines* article designed to dispel the notion that within quantum mechanics there is some essential hiding place for an unexplained consciousness phenomenon. We conclude that the conscious observer is *not* indispensable for reducing state vectors in

quantum measurement and that, for the most part, quantum mechanics is utterly irrelevant to philosophy of mind. The chapter finishes with a defence of the view on offer against some criticisms recently suggested by Nobel laureate Brian Josephson.

Armed with a plausible basic angle on the mind body problem and safe from a flanking attack from quantum physics, in Chapter 5 we outline another foundational position. Our approach here is to view the conscious subject as an abstract data structure instantiated by a material substrate and to view the conscious experience of that subject as changes to the data structure. We explore some of the ramifications of this "self model" data structure view for the way we understand conscious experience, and in the next chapter we see some of the low level representational and information theoretic properties we should expect the self model to exhibit. In the following chapter, our project is examining the relationship between the self model view and functionalism, and we see how some of the problems of the latter may be overcome by self models.

In Chapters 8 and 9, we place the self model view in the context of evolutionary biology and explore the neuronal group selection theory of Nobel laureate Gerald Edelman. Understanding Edelman's account of perceptual categorisation helps us to see what kinds of questions neuroscientific theories must answer, and it helps situate our own self model approach within a context of lower level cognitive neuroscience on the one hand and higher level subjective mental experience on the other.

Chapter 10 draws on examples from the European Symposium on Artificial Neural Networks to introduce the artificial neural network framework which we use in Chapter 12 to explore ways of instantiating materially the very basic properties of self models which we first outlined in Chapter 6. In the intermediate chapter, we comment on our use of the term 'information' and look briefly at how information may be processed by neural networks. Finally, we target the supposed incompatibility between the connectionist approach we have used and the classicist approach to artificial intelligence and cognitive science and argue that whatever criticisms the classicists may have of connectionism are not dangerous to our own project.

Chapter 13 roughly marks the beginning of the second half of the dissertation, both physically and conceptually. In the second half, our

analysis tends away from the basics of self models and toward some of the implications of understanding cognition through the dynamical systems instantiating them. A major task in this second half is also to elucidate the importance, if any, of chaotic behaviour in these dynamical systems.

We begin with the notion that dynamical properties of the low level instantiating material structure of a self model may have interesting effects on the higher level subjective experiences arising from it, and we inaugurate a new representational schema for discussing the relationships between these different levels. It is within this representational schema that, in the next chapter, we defend an argument for high level psychological indeterminism supervening on a deterministic low level material substrate. Since the line of thought was first suggested at the July 1993 conference of the European Society for Philosophy and Psychology, it has been the subject of an ongoing debate with Peter Smith, and this chapter includes the latest volleys in that exchange.

Next we turn to exploring one possible consequence of instantiating self models with an analogue material substrate such as neural wetware. The chapter is based on my suggestion at the Russian International Computer Systems and Applied Mathematics conference in 1993 that analogue systems which are specifically chaotic may pose a problem for Turing computability despite not violating the Shadowing Theorem. If it should turn out that chaotic analogue systems—such as the brain arguably may be—could behave noncomputably, and if higher level self model features rely essentially on such behaviour, then human cognition may not be modelled completely within the context of Turing computability (and the success of so-called "strong AI" will be correspondingly limited).

The next three chapters are more philosophy of science than of mind; here we engage Smith again over a series of arguments he has offered which threaten to reduce to irrelevance (in our cognitive science context) some of the points we have so far made about the rôle of any specifically chaotic dynamics in the wetware instantiating self models. First we defend our view of chaotic systems against the criticism that real physical systems are categorically unsuited to the infinitely intricate mathematical world of chaos theory. Next we clarify Smith's comments on the predictability of chaotic systems and discuss the recent discovery of physical models with so-called riddled attractor basins and their bearing both on Smith's view and on our findings in Chapter 15 regarding the

computability of chaotic analogue systems. (This recent discovery by applied mathematicians appears to vindicate those findings.) Finally, we offer a simplified approach to complexity which reveals some of the shortcomings of Smith's argument that the complexity of real chaotic systems may be essentially the same as the complexity of random noise. The chapter includes the definition of a new measure of complexity inspired by Bennett's pioneering work and intended to transcend the shortcomings of that earlier measure.

In Chapter 19, we return to the topic of representation and finish the analysis of complexity by observing some characteristics of its rôle in representational systems. We note that the complexity of a pattern is inextricably bound up with the representational system through which it is viewed and that what is complex from a third person point of view may be simple from the point of view of a system of which it is a part—and *vice versa*. We conclude that the rôle of chaotic dynamics in the material instantiation of a self model is for now an open question. In the last chapter, we finally conclude our discussion and review the positions we've taken, and the Appendix includes a statement on the authorship of this dissertation and covers the legal bases associated with reprinting previously published or forthcoming material.

1.4 What Isn't the Last Word—Return to Innocence

I have always suspected that at least something was wrong with any philosophy book I have ever read, just as with Taylor's *Metaphysics*, and this dissertation is a glimpse of the kind of view which has been simmering away for so long and making me think some other approaches were not *quite* right. I suspect that most of this book is not quite right either, but it is one step in sequence of ideas which are hopefully progressing.

For me, philosophy is and always has been a struggle to refine my world view, to grasp what Heidegger calls my *innerweltlichkeit*, to place myself in a context and to understand how my physical and mental existence relates to that of the rest of the cosmos. This dissertation is a portion of what makes up that understanding just now, a slice of what has come of a few years' proper research and a lifetime of philosophical wondering and wandering. I don't expect anyone will buy into the entire

picture on offer here—not even myself—but I will be satisfied if even a few titbits make their way from these pages into someone else's struggle to understand, that in another's struggle they may yet find more refinement than they have in mine.

2

What is it Like to be Nagel?

We start with a look at the distinction between the first person and third person descriptions of sensation in intelligent systems; this discussion will begin to establish the background for all our succeeding explorations of the relationship between these two points of view. The following is reprinted, with minor corrections and revisions, from Mulhauser (1993a).[1]

It is meant as a very light introduction to the brand of dual aspect monism which is central to much of what we will explore in coming chapters. Thus, it is by no means intended as a thorough analysis of the extensive collection of ideas which have grown around Nagel's views on subjectivity. For the interested reader, several papers in this literature, including Haksar (1981), Foss (1989), van Gulick (1985), and Pugmire (1989), are in various ways particularly compatible with the position we take here. Davis (1982a) and Flanagan (1985) are useful as attempts to move beyond Nagel's kind of approach, and Kekes (1977) and Malcolm (1988) attack Nagel's framing of the problem, while the sympathetic McCulloch (1988) offers an interesting line complementary to Nagel's own. We here make the initial case for a dual aspect monist approach and then explore and elaborate on the idea more carefully in the subsequent chapter.

2.1 The Subjectivist's Object

One objection to a physicalist account of cognition is the alleged incompatibility between subjective phenomenological descriptions of conscious "feeling" on the one hand and objective neurophysiological descriptions of things like brain states on the other. We might think of descriptions in roughly corresponding pairs, which are supposed to be of

---

[1] Please see the Appendix for information on all articles reprinted in this dissertation.

entirely different types: sensation/neural activity, mental event/physical event, mind/brain. In these pairs, the sort of description on the left is supposed to match "having a point of view" from the first person, while that on the right is supposed to correspond to the "scientific" perspective of the third person. The critic of physicalism argues, in effect, that the right hand can't know what the left hand is doing. That is, we can't know someone's point of view by looking at their physical description; we can't understand how an experience phenomenally *feels* in the terms of neurobiology.

Our immediate task is to suggest that in cashing out the subject/object distinction, we are left with dual complementary descriptions of the same thing—two sides of the same coin, as it were—and that both have a legitimate place in a comprehensive philosophy of mind. I argue that no theory could explain away the privileged status of the subjective point of view and that it should just be accepted as a necessary fact about the things we call "beings". In the first half of this section, we unpack some of the ideas behind the subject/object distinction and discuss what might be the obvious point that we may only know other beings from the third person. The second half introduces a thought experiment to help show why, given this conclusion, we as possessors of a subjective point of view still cannot help but accept that other beings *have* a subjective point of view. We start with a look at Nagel's (1974) famous question: "what is it like to be a bat?"

The question immediately tempts us to divide the world, rightly or wrongly, into what Hofstadter (1981) has later called "be-able things", or BATs, and non-BATs, or things to which it makes no sense to attribute a point of view. We might list dolphins, rats, philosophers, and of course bats in the first category and aqueducts, roundabouts, lumps of wax, and hats in the second. Nagel's point in the original paper was that the "subjective character of experience" cannot be captured by any reductive physical account of a thing. In the case of a bat, or any other nonhuman BAT, we humans can have no idea what it is like to be them because we can have only physical descriptions which *miss out* the subjective quality of their experience. Nagel is not after any physical description or behavioural observation, but rather the ineffable feel of what it is like *for the bat* to be a bat. Talk of enjoying hanging upside down, for instance, tells us nothing of what *bat* enjoyment of hanging upside down could be.

Nagel's argument is that I can't even *conceive* of bat phenomenology because it is too alien to anything *I* have ever phenomenally experienced.

But if we consider other non-bat BATs—namely humans—we can see that if anything, Nagel showed more than he said. What is it like to be Nagel? I propose that no answer to this question can get off the ground any better than any for the bat question. For even in the case where observable qualities of our experiences are similar, I still cannot fathom what it would be like to *be* Nagel. This doesn't mean it's not like *anything* to be Nagel, but it means I could not understand an answer to the question even if one could be given. Pretending, for instance, that I have some particular memory which Nagel reports (pretending that I remember going spelunking, for instance) is clearly not the same as *having* the memory for myself (actually *remembering* going spelunking). Even knowing something of Nagel's habits of thought or his heuristics for solving certain kinds of problems is not the same as actually *having* his habits of thought or *using* his heuristics.

Nagel tries to steer clear of the issue of communicating what it is like for something to be itself, but my point is that Nagel's question could have no other point. Let's look more closely at what it could mean for there to be an intelligible answer to the question of what it is like to be Nagel. Consider a hypothetical case in which I empathetically took on *as my own* every single memory, habit of thought, and so on of the famous philosopher. This comes much closer to *being* Nagel. Of course, I would also need all Nagel's physical characteristics so I would know the kinaesthetic feel of his physiology, and I would have to forget all my own memories and habits of thought. In what (if any) sense, then, would I still be *me*? I couldn't *know* I was me, because then I would know I wasn't *Nagel*. But of course I *would* know I was *me*; I would be *Nagel*. The entire notion wreaks havoc with ideas of self identity. While I was being Nagel, where would ex-me be, and would I have two simultaneous points of view, an old one and a temporary new one? What would be the point of view of ex-Nagel, whilst I was borrowing his—the view from nowhere? One is reminded of the Taoist who awoke from dreaming he was a butterfly. He wondered to himself: am I a man who has just dreamt he was a butterfly, or am I a butterfly who even now dreams he is a man? (Graham 1981, p. 61)

Nagel calls the ability to imagine being in the conscious state of another being *sympathetic imagination*. What the above muddle suggests is that we simply could not sympathetically imagine *being* someone else. Nagel notes in passing that a kind of solipsism results if we confuse sympathetic imagination with what he calls *perceptual imagination*, or the capacity to put oneself in a state resembling the state one is in when one actually perceives a thing. If we misinterpret sympathetic imagination as if it worked like perceptual imagination, Nagel correctly points out, it would seem impossible to imagine anyone else's experience but our own. But Nagel neglects his own point: his argument suggests that I might imagine what it would be like for *me* to have someone else's experience, but I couldn't imagine what it would be like for me to be *them* having the experience. As much as I might imagine what it would be like for me to experience Nagel things or bat things or any other BAT things, I couldn't imagine what it would be like for me to *be* Nagel or a bat or any other BAT.

Although Nagel explicitly rejects the idea, perhaps what we are really after is some sort of *translation* from Nagel's experiences into the kinds of experiences I can understand (namely, my own) rather than actual knowledge of *being* Nagel. We might consider someone's explanation along the following lines: well, when Nagel gets up in the morning, he feels *like* you do except that he starts thinking of scones where you would think of muesli and toast, and scones taste to him *like* croissants do to you...and so on. But what if there were experiences (and there undoubtedly would be) for which I had no analogue? Moreover, who could, even in principle, supply such a translation? What would it mean for *that* person—who would know what it was like to be both of us separately—to know what it is like to *be* either of us?

None of this requires that we be sceptical about it's meaning *something* for other beings to be themselves; but we are faced with the implication that we cannot have a public language description which could make intelligible what it is like to be a thing being itself. To put it tritely: we can't have a description and be it, too. That is, we can't be offered a description and infer from it what it is like to be the object of that description. Perhaps Nagel anticipates this when he says that we only infer that other humans, for instance, have conscious experience. But what kind of inference could this be? On the face of it it is not inductive,

because no one is possessed of a sort of "phenomenon detector" which could establish that all similar humans examined thus far have had conscious experience. It does not seem deductive, either, since Nagel himself denies the appearance of any physically or logically necessary connection between observable behaviour or brain states and conscious experience.

The view we shall explore here is that we cannot help but infer that other humans have phenomenal experience on the basis of their similarity to us, despite the fact that we cannot describe what it is like for them to have it and despite Nagel's denial that there is any necessary connection between brain states and mental states. The strategy of avoiding a kind of phenomenological solipsism we shall call *cybernetic realism*. (We adopt Wiener's term "cybernetic" to indicate that we are concerned here not just with human BATs; but as it will emerge in subsequent chapters, we are not committed to attributing experience to anything close to Wiener's broad class of cybernetic systems.) Similar to other brands of realism, cybernetic realism is the principle that there *are* (objective!) matters of fact about the subjective experience of other BATs, independently of our actual or potential state of *knowledge* of those facts. Acknowledging that we cannot know anything about another BAT's subjective experience doesn't mean there are no BATs. Despite it's superficial similarity to Dennett's "intentional stance" (1987), cybernetic realism asserts the reality of the subjective quality of experience, what some call the qualia—that which Dennett explicitly wants to leave out (1988).

So far, then, we've established the unremarkable fact that the first person point of view applies only to the first person. Nagel was correct in maintaining that any physical description of a system must completely miss out the answer to the question of what it is like to *be* the system. But the point is stronger: no description *at all* (physical or otherwise) could by itself make intelligible what it is like to *be* another BAT, full stop. Of course this is *not* to undermine the case for any kind of sympathetic communication at all; it is only to say that some other ingredient is required in addition to the description itself.

Is the physicalist, then, doomed? The physicalist is usually cast as aiming for a complete reduction of all mental-speak to physics-speak. If physicalism cannot provide a physical description which is provably

identical to a given mental state (or from which the qualities of a mental state may be logically inferred), has it failed its mission? The answer is no. The domain of physicalism is the physical, the objective; we should not ask that the physicalist—or anyone else—tell us how a given physical description feels for the being doing whatever is being described, but only what objective physical structures lie at the heart—or brain—of being that doing being. Physicalism should give neurological descriptions of the structures which enable particular cognitive tasks to be performed but not of how those cognitive tasks feel when performed with those structures. Moreover, no theory could in principle give us an account of how *by virtue of* such and such a physical structure, this system is having such and such an experience with this or that phenomenal quality. We turn now to a thought experiment which may show why cybernetic realism together with a physicalist account of the objective aspects of cognition should be palatable to any philosopher of the mind, regardless of—or perhaps because of—her own experience with a unique subjective point of view on what it means to *be*.

2.2 The Objectivist Objects—Science to the Rescue

Suppose a lecturer from a far away land comes to visit a British university and brings with her a strange device she calls a lanemonehp mirror into which she claims she can look and, by studying someone's reflection, see secret properties of their brains and tell whether they are having phenomenal experience. The device supposedly cannot be fooled by anyone's normally observable behaviour (internal or external) but works only by sussing out secret internal properties. The lecturer has made a name for herself as the "which-is-which" doctor (or "which" doctor, for short) because of her claimed ability, with the aid of her lanemonehp mirror, to tell exactly which brains are producing phenomenal experience and which are not. She travels the world, assisting psychologists with their experiments and delivering lectures to Philosophy and Cognitive Science departments.

Like Santa Claus, who knows just who has been bad and good at holiday time, the visiting "which" doctor makes a habit of announcing to her class at the end of each lecture exactly which students have failed to have conscious experiences during that day's talk. Her students make a

sort of game out of trying to guess in advance which of their classmates are having phenomenal experiences and which only appear to be having them. They have learnt from the "which" doctor that even some of those unlikely candidates who appear externally to be comatose are still having conscious experiences. Likewise, a few students find that some of their best friends are always, during lectures at least, devoid of phenomenal experience.

What should we make of the "which" doctor's claims? Her alleged ability amounts to being able to select from a group of people with similar internal and external behaviour just those who have what philosophers call *absent qualia*. Philosophers call such creatures *zombies*, and many claim that there is nothing logically inconsistent about imagining a zombie who has perfectly normal neuronal and external behaviour but who nonetheless has no phenomenal experience whatsoever. If the philosophers' claim is true, oughtn't we be suspicious of the lecturer and her device? In particular, given the impossibility of knowing what it would be like to *be* someone else (and thus, probably, whether their *being* is *conscious being*), how could the "which" doctor's lanemonehp mirror work?

Now suppose the doctor lets us in on her secret: it turns out that her special mirror doesn't show any secret properties of brains at all; it merely allows her to examine in great detail the overall functional arrangement and activation patterns of a given brain. Her special talent has nothing to do with hunting missing qualia—indeed, she steadfastly insists that qualia go missing only when something has gone awry in their neurophysiological homes. According to her, there *is* some kind of limited logical connection between brain functioning and mental experience. She says she is only examining phenomenal experience from the reverse point of view to her subjects—the mirror image—and seeing from the third person the objective neurological properties which correspond to the subjective experiences of the first person.

When pressed for the details of how she could know about conscious experience on the basis of the functional characteristics of neural arrangements, she explains that after developing the technical pieces which allowed her lanemonehp mirror to show neural behaviour in such great detail, she used it first to examine herself. By subjecting herself to a wide range of experiences while reflecting, as it were, on the patterns of

neurophysiological responses in her brain associated with various kinds of experiences, she discovered remarkable trends. In hundreds of trials, she found the same kinds of neuronal responses always correlated with the same kinds of phenomenal experiences. Eventually, she even took films of the mirror's image while colleagues administered to her drugs which induced episodes of unconsciousness or other impairments to her normal mental life. In these cases, she discovered similarly robust correlations between the lack of conscious experience on her part and fundamental changes in the interactions between her neurons.

Eventually, she made similar tests with other subjects and uncovered every time the same kinds of correlations between observed neurophysiological activity and at least their *reports* of their experience. Because she couldn't conceive of *her* qualia going missing while her internal and external behaviour remained the same, she reasoned that qualia weren't just in her head. By virtue of the similarity to her own of their internal and external objective properties, she came to believe—although she couldn't properly know—that the other subjects were keeping qualia, too. It was her first step in becoming a cybernetic realist. The argument which was the foundation of her belief extrapolates[2] to other BAT type animals whose brains might function similarly under the light of her lanemonehp mirror. Arguments of the form are the logical underpinnings of cybernetic realism.

2.3 Does it Matter What it's Like?

We've established, then, that Nagel's question about bat-hood was even more significant for the distinction between subject and object than he indicated. We've seen it is the very strength of the distinction which renders it ineffective as a criticism of physicalist accounts of mind—because the subjective point of view is so firmly anchored to the first person, there *could be* no account of it, physicalist or otherwise. The subjective and objective appear as two distinct aspects of the same thing, and the subjective aspect is that which, like the image in an ordinary (phenomenal, not lanemonehp) mirror, reflects our every objective change, but which we still cannot approach objectively no matter how earnestly we reach out for the mirror's image. We have seen reason to

[2] We see in the chapter following some reasons we might believe this extrapolation.

believe that internal and external behaviour which is observably similar to our own is a sufficient but not necessary condition for a thing's being a BAT. We should, in short, give up on zombies and other supposed problems and get on with answering what questions we can in the style of the "which" doctor and physical theories of mind.

In the next chapter, we expand on these ideas and come to a better understanding of the brand of dual aspect monism which can answer the kinds of difficulties to which Nagel's work seemed to point.

3

A Rose, By Any Other Name

To be sure, Nagel's "what is it like..." argument doesn't exhaust the set of approaches to questions about the alleged incommensurability between first person subjective phenomenological descriptions and the third person objective accounts of neuroscience.[3] Indeed, there may be considerable problems just trying to imagine what it would feel like for *ourselves* to have an experience described in objective neuroscientific terms,[4] let alone trying to imagine what such an experience would feel like from the point of view of someone we are not. If we are to establish the plausibility of a kind of dual aspect monism to underlie our coming discussions about how to locate conscious experience in a material substrate, we must address other aspects of the claimed incommensurability between phenomenology and the objective accounts of the third person.

In what follows, we examine more closely the ideas which lead to this claim of incommensurability and explore some ways in which they may be confused or simply incorrect. We begin with the general notion that from an account of the physical properties of an object, we cannot understand what it is like as an object of sensation. The particular example is concerned with understanding from its physical description how a rose phenomenally smells. Building on a number of observations about this case, we then move on to analyse similar examples about neurological descriptions of experience both in ourselves (the first person case) and in others (the third person case). The conclusion is an expansion of the idea in the second chapter that correlations between neural activity

---

[3] The present material also bears on Kripke's (1971) and Jackson's (1982) related arguments and on Searle's (1992) summary (pp. 117-118) of these two kinds of positions together.

[4] That is, it's often easy to imagine ourselves having an experience described in subjective phenomenological terms—imagining smelling a rose, for instance—but not so easy to imagine such an experience described in objective neuroscientific terms—imagining having a pattern of activity across the olfactory cortex, for instance.

and sensation as well as the organisational patterns at the neural level which may enable phenomenal sensation in the first place are legitimate areas for systematic scientific inquiry and that claims of incommensurability are singularly unhelpful.

3.1 Smelling As Sweet

Suppose for the moment that you have never before smelled a rose. Maybe you've seen them on greeting cards or on the television, and maybe you've heard romantic stories about passionate young men or women exchanging them with the people they love. You might even have books about botany and chemistry which tell you the structural and chemical properties of different kinds of roses. Maybe you have dedicated your life to understanding the physically describable properties of the flower, yet somehow in all your studies you have neglected that one single thing: you haven't a clue how a rose smells. In a moment, I will suggest that the fact you can know all the properties of a rose but be ignorant of its smell is nothing more than an accident of the neural structure of *Homo sapiens* and that, had this neural make-up been slightly different, you could easily have figured out how a rose smells just on the basis of all this objective physical data about the flower. Moreover, I will suggest there is a particular way in which even someone with *Homo sapiens* neurophysiology *could* know the smell of a rose just on the basis of its physical properties.

But first, let's note some of the implications which appear to follow from the fact that on the one hand we could know all the physical properties of the rose but on the other hand we might still not know how it smells. On the face of it, it seems obvious that the smell of a rose cannot be described in terms of the flower's physical properties. That is, a description of the phenomenological quality of a smell sensation cannot be logically deduced from a physical description of the object of that smell sensation. There might be many ways to come to the knowledge of how a rose smells, but learning all the physical data about a rose does not look to be one of them. It is not a straightforward inference from this position, but it does seem at least a plausible extrapolation from this position to the notion that no phenomenological description of *any* sensation can be derived from a physical description of the object of that sensation. There

is a similar but weaker extrapolation which some might like to make to the position that no phenomenological description of a sensation can be derived from a physical description of the neurophysiological properties of a system claiming to be having that sensation or, similarly, that no phenomenological description of a sensation can be derived from a physical description of the neurophysiological properties *we* might exhibit if we actually *were experiencing* that sensation. But with these possible implications in mind, let's for the moment return to the briefly stated claim above that it is just a neurophysiological accident that we cannot know how a rose smells just by learning its physical properties and that, in fact, there is a special way to accomplish just that.

The justification for this claim actually derives directly from the observation that there is after all a way to smell a rose just by learning its physical properties. Notice that some members of the set of propositions describing physical rose properties describe the physical effects of rose chemicals on human nerve cells. Granted, this goes beyond the kinds of properties of roses which normally come to mind such as colour, anatomical structure, and molecular structure of various constituent chemicals. Yet, on the face of it, anyway, there is nothing to prevent our allowing this set of physical properties of a rose to include its physical effects on a human nose (or on dead skin cells, or lumps of wax, or ski mountains, or cloves of garlic, or any other physical thing). But even so, says the proponent of the incommensurability claim, knowing how a rose affects nerve cells is no help in knowing how it feels to be a human being whose nerve cells are responding to a rose and who is actually smelling a rose.

3.1.1 Smelling a Rose in Three Easy Steps

Let's make a brief note about the word 'information'; we'll return to the topic briefly in Chapter 4 and again in Chapter 11, but it is important for the moment to clarify the word before we continue. In particular, our application of the word 'information' is based strictly on correlations or lack thereof between the states of physical entities. Thus we may say of someone who speaks English and only English that "reading" a sentence in French still conveys information. In our use of the word, what is printed on the page still conveys information even if the reader only recognises characters and word endings and so forth (or even just lines

and curves). Now, if our English speaker *also* understands French, then reading the same sentence conveys information in the manner of the original example and *more* by virtue of the correlations between the words read and words the reader has already learnt to recognise. Just to anticipate the forthcoming discussion a little, it is useful also to notice that the same physical object may pass on information in more than one modality and at more than one level: the same French newspaper may be read or smelt or tasted, and the same newspaper may again be perused by an English-only speaker (who takes in information at a low level, by recognising characters) or by a French speaker (who takes in information also at a higher level, by recognising entire words and their meanings).

With these comments in mind, we can see that there are many different ways to take in the physical information about roses and especially about their effects on human nerve cells (and we can address the point which ended the previous section). We might read the information printed in books in Roman characters, or we might hear it in lectures or when someone reads a book to us, or we might feel it in books printed in Braille. Or, particularly in the case of the effects of roses on nerve cells, instead of taking in the information at such a high level with sight or hearing or touch, we might artificially stimulate the nerves leading to our olfactory cortex, or we might even stick a rose under our nose and let *it* stimulate nerves leading to our olfactory cortex. We might even read information about how the rose excites human noses and then use that information manually to artificially stimulate our noses! We are still taking in *physical* data, but those data are entering our system at a different level than with simple reading. (See Marr 1982, Horgan and Tienson 1993, and Chapter 13 for more on levels.) Instead of exciting nerve cells in our retina (when we read the data) or nerves whose efferent signals go to the auditory cortex (when we hear the data) or some such, they are exciting nerve cells whose outputs go to the part of our brain which processes smells. So perhaps there is a way, after all, to know the smell of a rose just on the basis of its physical properties, as long as the information about those physical properties is made available at a level where it can do some good.

This way of approaching the problem is analogous to what I might observe if I typed into my computer a book all about the way my computer's disk drive works. I might include detailed specifications of the

exact electrical signals which activate the disk drive to perform particular functions. I might even include some specific examples of the signals which would be sent inside the machine to, for instance, format a disk. Yet no matter how I typed into the machine these descriptions of the computer's internal signals, I could never get it to format a disk. Descriptions of chip-level disk interface signals have no effect whatsoever on my computer's actual chip-level disk interface signals, and they cannot make it format a disk. Yet, if I got into the machine's insides and found the appropriate control lines and imposed my own signals on them, I could cause the machine's disk drive to do just that. Likewise, if I selected an appropriate operating system level option to format a disk—which in this case is analogous to shoving a rose under my nose—presto! the disk gets formatted, on the one hand, or I get to smell a rose, on the other. Yet just because I cannot activate my computer's disk drive by typing into it a detailed description of the internal electrical signals which operate it does not mean that such a detailed description "misses out" some important feature of the way the disk drive works.

Now, it is purely an (intentional!) accident of the design of the computer that I cannot change it's internal chip-level signals by typing into it a description of what I would like those chip-level signals to be. But we can easily imagine a computer which did allow direct manipulation of some subgroup of its chip set by simply typing in a description of what signals chips in that subgroup should send. Likewise, it is purely an accident of neurophysiology (an "accident" which occurred, no doubt, for good evolutionary reasons) that we cannot change the output of nerve cells leading to the olfactory cortex by reading or hearing or touching a description of how those cells *would* fire if there were rose chemicals directly stimulating them.[5] That is, there is no neurophysiological reason why we couldn't, in principle, stimulate our own nerve cells where olfactory information is processed in a fashion similar to the way we stimulate our own nerve cells which activate our various muscle groups.

---

[5] In a way, the computer example is slightly misleading, because selecting an operating system level option for "format this disk" is still possible from the keyboard, whereas apparently there is nothing we can do by reading or listening or touching which gives us the effect of smelling a rose. But the disanalogy can be remedied just by supposing the computer has a special "format button" or "format switch" instead of a keyboard command, such that we could type anything we wanted and still not get the drive to format a disk without pressing the special switch.

(Indeed, as it will emerge later in our discussions of self models, I believe something like this is the basis of imagination.)

At this point two important notes of logic demand clarification. First, we are not yet arguing for a particular position on the mind body question. The present picture makes sense under any view which allows that a physical interaction with something is the cause of a sensation. That is, the pin-prick causes the pain, the empty stomach causes the hunger, the Union Jack causes the sensation of red and white and blue. What happens *after* the initial physical interaction can, for the moment, be almost anything: there could be a bunch of neural firings culminating in a Cartesian communication with the mind realm, or there could be some of Sir John Eccles's psychons collapsing wavepackets at synaptic junctions, or there could be a little conscious homunculus sitting inside a head watching neural signals appear on tiny monitoring screens. Secondly, the present position is no argument against anyone who supposes that there is some extra property or spirit which somehow gets passed on from rose to neuron or that there is more to the interaction of a rose and a single neuron than can be described in physical terms (although probably other damning arguments could be mustered against such a position, and it is hard to imagine what non-physical element there could be to such an interaction, especially given that mind predicates, for instance, don't normally get applied to the likes of single neurons or olfactory cortexes). The point is only that *if* we allow an exhaustive description of how a rose affects a single neuron into the set of propositions about physical properties of roses—which seems perfectly straightforward—and *if* we believe physical interactions are the cause of sensations, *then* we must question the logical force of the claimed incommensurability between objective physical descriptions and subjective phenomenological ones.

Strictly speaking, denying the incommensurability claim does not amount to asserting the claim that phenomenological descriptions can be logically *derived* from physical descriptions. Under the usage adopted here, information is conveyed by correlations between the states of physical entities, and these correlations are typically established causally. Thus, a description of rose chemicals may not logically entail a rose smell, but that doesn't imply that rose chemicals or even just a thorough

description of them may not pass on the information required to experience the smell.

To use a related example: if you are offered a physical interaction (information) in the form of a pin-prick, you cannot logically infer that you are about to feel pain, but given the pin-prick, you can—at least in part because of the way the nervous systems of human beings are structured—come to know that it hurts. The connection here is empirical, not logical—indeed, it is causal—but few people experienced with pin-pricks would deny that there is rational force to the proposition, "if I prick you with this pin, then it's going to hurt". Likewise in the case of the physical description of the rose, except that here we must be more careful in characterising the connection. Given appropriate use of physical information about the rose, we may come to know how it smells. (And given a different neurophysiological make-up, even with single-modality access to physical information through Braille reading, for instance, we might still come to know how it smells.) If the rose information is made available directly to the appropriate modality (by sticking the flower under the nose), this interaction initiates the same kind of causal chain as the pin-prick. If it is made available to a different modality (enabling us to undertake artificial stimulation of nerve cells or, if we were "wired" differently, enabling us to provide these data to the right nerve cells), then it facilitates our own initiation of such a causal chain in the same fashion that someone offering us a pin facilitates our pricking ourselves and feeling pain. In this case, few people experienced either with using an artificial apparatus to stimulate their nerves or with shoving flowers under their noses or even (if they had odd neurophysiology) with making cross-modal use of information at an appropriate level would deny that there is rational force to the proposition, "if I give you this information about the physical properties of the rose, then I'll enable you to work out how it smells".

3.1.2 A Tasty and Melodious Side Note

The tentative conclusion that no phenomenological description of a smelling sensation could be derived from an objective physical description of the object of that sensation is, on this account, more a matter of physiology than of logic. Indeed, there is good reason to believe that if humans were wired in such a way that read descriptions of rose chemical

properties *automatically* brought about the appropriate kinds of firings in the olfactory cortex (perhaps by way of some high level linguistic to low level olfactory "short circuit"), then humans might well believe that rose smell is *logically entailed* by descriptions of rose chemicals. It is worth reflecting on the degree to which our natural understanding of P from (P & Q), for instance, might be predicated on the neural circuitry of our brains. It would not be at all surprising to discover some defect in some humans whereby they were incapable of recognising P from (P & Q), and for such humans P would clearly *not* follow obviously from (P & Q). It might be objected that such humans don't properly *understand* correct usage of a sentence like P or a sentence like (P & Q), but it is arguably the case that the reason everyone else does understand the sentences is a result of their having "normal" neural circuitry as compared to the "abnormal" circuitry of those who don't understand.

In fact, rather than wondering about humans who might fail to recognise something as "simple" as the P in (P & Q), we might instead get at the question by considering some whose associative connections are a superset rather than a subset of these kinds of logical connectives. For about ten adults in one million, sensory input in one modality regularly, uniformly, and involuntarily invokes sensation in a second modality. For someone with *synaesthesia* (Cytowic 1989, 1993 for an introduction; see also the rare 1992 book by psychedelic theorist Terence McKenna and Timothy Ely), the word 'logic' might be a deep blue colour, and for another, a middle C might taste slightly of grape.[6] The phenomenon has fascinating implications for creativity and perceived metaphor (O'Malley 1964; Ralston 1976; Marks, et al 1987), and it suggests limitations in standard views of how blind people might understand colour (Wheeler 1920, Wheeler and Cutsforth 1922). (Note that positron emission tomography reveals cross-cortical blood flow, indicating activity in the visual cortex, for instance, even when a synaesthetic subject is blindfolded and her primary stimulus is auditory.)

Suppose now that synaesthetes were the norm and "normal" people the exception: then the colour orange might be just as obvious an

---

[6] Indeed, I had until very recently assumed most people would agree with me that the number 5 is scratchy and the number 9 sharp, while 4 is soft and comfortable, almost fluffy. My own experience with such associations, however, is nowhere near as vivid as that described by people like Michael Watson (discussed by Cytowic), who feels shapes when he tastes.

implication of (P & Q) as most of us consider P to be, and those who couldn't "see" the orange in it would be thought of as just as impoverished as we might consider those who don't see the P. (In fact, while individual synaesthetes are remarkably self consistent in the associations they report, there is rarely agreement in the associations reported by *different* synaesthetic subjects; thus, our example requires a small cheat to ensure that the hypothetical synaesthetic norm is uniform.) Indeed, we might think of the P from (P & Q) connection as a kind of "single modality synaesthesia" with which we are almost all affected. Perhaps, after all, the great foundation of logic on which analytic philosophy is built is itself empirical, a product of the conceptual connections we involuntarily experience when we read a sentence like (P & Q) but nothing more. Probably most "normal" people will insist still that there is something about seeing the P in (P & Q) which makes it different to the synaesthete's experience (apart from being in a single modality), and we won't here belabour the discussion, but it is no easy task to pin down exactly what it might be that makes logic seem *necessary* apart from arguably *contingent* perceived connections which might perhaps be traced merely to the way "normal" brains are put together.

There are titillating implications here from the idea that the *rational* force of connections between sets of propositions may stand or fall according to *empirical* facts about human brains and the *empirical* uses to which physical information may be put, but our point in exploring the issue has been merely to suggest that any suspected impossibility of experiencing a smell sensation, for instance, on the basis of a written description is chimeric.[7] For now, we shall leave the exploration of the other interesting implications for another time and return to the remainder of the tentative extrapolations from incommensurability which we noted above.

---

[7] Of course, in the case of synaesthetes, it is the words *themselves* which have smells and not descriptions, but our point is merely to note the neural plausibility of cross modal induction of a sensation. There is every reason to believe that if synaesthesia is possible, then it would also be possible for words or combinations of words (like descriptions) to be *systematically* correlated in such a way that the induced sensations matched those which would be caused by the object of description.

3.2 Smells and Other Good Sensations

First was the extrapolation from the idea that no phenomenological description of a smell sensation could be derived from an objective physical description of the object of that smell sensation to the idea that no phenomenological description of *any* sensation could be derived from an objective physical description of the object of that sensation. We have not taken any great care to define 'sensation', and there may be some use just in keeping it as an undefined term, but for the moment let's restrict ourselves to whatever brand of such things may be the immediate result of sensory interaction with the physical environment. That is, I mean the pain feeling and visual quality of encountering a flying mallet, but neither the physical damage of such an encounter nor the abstract perceptual categorisation of a visual pattern as a mallet;[8] and I mean the bewilderment of seeing a flying mallet but not the feeling of anxiety or paranoia of wondering if someone might in the future throw a mallet at you. At least when we restrict ourselves to these kinds of sensations, then, the same kind of argument we've applied above applies here: it is a mere accident of human neurophysiological design that we are incapable of bringing it about that the relevant nerve cells are stimulated except by actually experiencing the event in such a way that those nerve cells get stimulated by the environment. That is, we can't get the afferent signals in any way save by actually having someone throw a mallet at us. If we could—for instance, if we had special "mallet nerves" which could excite other nerve cells in just the same way a mallet would, or if we had some variety of mallet-specific synaesthesia—then there doesn't seem to be any reason to think that we wouldn't experience just the same feeling (but not the same physical damage) as actually encountering a flying mallet. (Again, this is unless we believe there is some special non-physical element of mallet-nerve interactions.)

The other two extrapolations described above fall to a similar kind of analysis. I suggested first that someone might like to argue that no phenomenological description of a sensation can be derived[9] from a

---

[8] This is true except insofar as perceptual categorisation contributes to the visual quality of the mallet; I take it that neither implies the other.

[9] As before, we are concerned with what *information* may be passed on to a specifically conscious audience by a physical description, which is arguably a superset of the set of propositions which may be logically *deduced* from that same physical description.

physical description of the neurophysiological properties of a system claiming to be having that sensation. In the sequel, we shall refer to this as the *third person problem*. The *first person problem* is the similar position that no phenomenological description of a sensation can be derived from a physical description of the neurophysiological properties *we* might exhibit if we actually *were experiencing* that sensation.

3.3 What's a Little Description Between Friends?

First we must be clear on one key to unravelling both these positions: sensations belong to conscious beings—to selves—and not to such things as descriptions, say, written in a book. Thus, by a physical description, we mean a physical description *as* read or felt or heard (or whatever) by some conscious self, who may be, for the sake of simplicity, just an ordinary human being. The suggestion is not that a book of the complete physical properties of a rose smells like a rose, or some such absurdity. We mean merely that such a book may capture perfectly well all the physical information there is about the rose[10] but that until that book is read—and appropriate use is made of the data therein—there is no rose smell to be had. This is no more mysterious than saying of sound—*qua* a sensation of conscious selves who possess the appropriate hearing apparatus—that there is no such thing when a tree falls in the forest and there is no one around to hear it. It is no more mysterious than saying no one has learnt the story of *Alice in Wonderland* until they have read the book or than saying a photograph of my mother contains all the colour information about how she looked on a particular occasion, even when the photograph is in a completely dark room where there is no colour to be found.

For anyone still not convinced, consider what it would mean if there *were* some other additional special property which somehow could be added to a description which would then make anyone immediately sense the object being described. Would it be any kind of objection that an ordinary video camera, when pointed at this new fortified description, isn't compelled to have the same kind of sensation we do? Likewise, would it be any kind of objection if we found that the same video camera

---

[10] Under our use of 'information', this suggests that the book represents physical states of the rose in a systematic way; see also Chapter 11.

couldn't smell the rose we've shoved first under our nose and then under its lens? Of course it wouldn't be an objection at all. Whenever we concern ourselves with questions about physical descriptions and what may or may not be missed out by them, we mean descriptions *as* interpreted by someone or something for which it is possible that a description may have missed something out.

3.4 The First Person Problem—Whichy Mirror on the Wall

Now, let's examine the second of the two positions described above, what we have called the first person problem. The lynchpin of our analysis of learning how a rose smells on the basis of information about the physical properties of roses was the observation that information must be made available at the right level, to a part of a system (such as a human) capable of using that information in the right modality. This is again crucial in understanding the first person problem: just as the book must be read to understand the story—and once it is read, in the right language and so on, it may be understood—so, too, must the description of my neural behaviour allegedly correlated with my having a particular sensation be experienced in the correct modality in order to be convincing.

Let's suppose I am equipped with the cybernetic realist's lanemonehp mirror of the previous chapter. Let's say I have collected a huge database of different kinds of sensations and my corresponding neurophysiological activity while I experienced the sensations. My database is so large that for almost any sensation I can ponder, I have data which indicate what kind of neurophysiological activity is generally occurring during the sensation. If I were the betting sort, I would probably put my money on the proposition that if anyone offered me a description of one of my possible patterns of neurophysiological behaviour, and the pattern offered matched one in my database, I could guess (on the basis of the recorded correlations) what my sensation would be like if my brain were exhibiting that pattern of behaviour.

I might have no deductive *logical* grounds whatsoever for believing I could know what my sensation would be if my brain were exhibiting a particular pattern of behaviour, but I would still feel justified in betting because of the data I had collected previously. Likewise, I might have no *logical* grounds for believing the sun will rise tomorrow, but I would still

feel it a pretty safe bet if someone wanted to bet it wouldn't. Just like the "which" doctor of the previous section, on the basis of experience I might find it inconceivable that my neurophysiological behaviour could be the same as that which in the past correlated with a given sensation and yet I could feel a different sensation. (Of course we mean "inconceivable" not in the sense that it is logically impossible, but in the sense that it is inconceivable that, *ceteris paribus*, being pricked with a pin right now would not cause me pain.) Even so, there might still be no deductive logical reason to deny that tomorrow the neural reactions caused by encountering a flying mallet suddenly might become strangely pleasurable, or might make me hungry, or might make me feel like doing arithmetic or going skiing.[11]

Just as in the previous section, I would suggest that indeed there is no logically necessary connection to be inferred from the kind of database of "mere" correlations I described above. Yet, I also suggest that these correlations are reason to suppose not some kind of putative cause and effect relationship between neural behaviour and phenomenal experience but instead to suppose that the phenomenal experience *just is* the subjective aspect of the neural behaviour. That is, by the phrase "pain caused by a flying mallet", we mean not some kind of cause-effect relationship between the firing of nerves in my head and a resultant (perhaps epiphenomenal?) sensation of pain; nor do we mean the "pain *just is* this firing of nerve cells in response to a flying mallet"; we mean instead something along the lines of "the pain *just is what it is like* when nerve cells are firing thus and so in response to the impact of a flying mallet". That this subjective aspect of our own objective neural behaviour is "real" is analogous to the fact that a cause of an effect or an effect of a cause is "real": we cannot infer it, we can only observe it...over and over again.

3.5 The Third Person Problem—A Tale of Two Arguments

Let's turn now to the most difficult question, that of the third person problem: the position that no phenomenological description of a

---

[11] Our later discussions of self models, however, suggest that such a change taking place would require a change in the functional rôle of those neural reactions to flying mallets and a subsequent alteration to the dynamics of the data structure instantiated by the brain of which the relevant neurons are a part.

sensation can be derived from a physical description of the neurophysiological properties of another system (such as another person) claiming to be having that sensation. (Put this way, the third person problem is closely related to the host of problems often grouped under the heading "problem of other minds".) With respect to the first person problem, we explored the case for at least *believing* that specific neural activity patterns in my own body would feel like particular sensations by appealing to the possibility of establishing a large database of correlations between previous sensations and the simultaneous neural activity in my own body. Now, with respect to the third person problem, we can first observe that again there almost certainly could be no *logical* inference from observed neural activity to phenomenal sensation. And insofar as I cannot be anyone save who I am, and I cannot logically infer what the sensation of another individual feels like *for that other individual* (the point of the previous chapter), I cannot build a database of correlations between *their* neural activity and *my* experienced sensations.

Moreover, in the first person case, the difference in neural activity between, say, seeing red and seeing blue, might be localised in a small area of the visual cortex; but in the third person case, while differences in neural activity between *that third person's* sensation of red and blue may be relatively localised, differences in neural activity between *my* sensations compared to the *other person's* sensations might involve much larger areas. If our two systems were coincidentally wired in just exactly the same way, a case might be made (without necessarily taking on board any kind of type-type or type-token identity theory) for suggesting that the other's experiences were just like mine, but the more different they are, the less plausible such a case may be.[12]

We have made much of the point that information about objective physical properties can be introduced into a system in a number of

---

[12] Since the human brain contains a number of neurons on the same order of magnitude as the number of stars in the Milky Way Galaxy, the idea of two brains physically being wired "identically", to within even a moderate degree of error, is absurd. We might, however, view the footnoted sentence as a variation on the correlation database example for the first person problem. We might say that our own systems at different times constitute different systems with very similar structures and that the reliability of past correlations between states of these different systems and experienced sensation is the basis of an inductive inference that future patterns of neural activity similar to those in the past will phenomenally feel the same as they have in the past. But to use an analogous argument to address the third person problem is, I think, cheating, on the grounds that it may beg important questions about identity over time.

different ways and at a number of different levels. Here again, the point applies, but in a more problematic way. In the third person case, not only can we not build up a database about neural activity and the corresponding sensations (although we could build up a database of corresponding *reported* sensations), but for the most part we cannot make information about the objective physical properties of another system available to the appropriate level of our own systems. As we alluded to above, we cannot take on all the necessary neural characteristics of another individual in order to see what sensations they are experiencing. If all we have to go on, even in the best case, is the kind of set of correlations we posited in the analysis of the first person problem, then it might seem there is no way to "get out" of our first person and say anything at all about the sensations of others on the basis of their neurophysiological characteristics. But all is not lost.

3.5.1 On a Bad Argument from Science

Against those, such as myself, who would argue that it is legitimate to extrapolate from characteristics of our own sensations to characteristics of another's sensations (or, for that matter, from the existence of our own minds to the existence of other minds), it is often objected that we are never permitted in science to extrapolate from the observation of a single case to all similar but unobserved cases. It is only after many observations of many independent cases that we are allowed to conclude anything at all. But this is not strictly true, and the sense in which it is not strictly true is just the sense we need to rescue from scientific incredulity the extrapolation I would like to make.

In scientific investigation, we expect the explanation of any single occurrence of an event to be consistent with the explanations for all other previously observed events, whether or not these past events are directly related to the event in question. For instance, insofar as we have confidence in the relative correctness of quantum electrodynamics (i.e., the congruity between whatever predictions we might make with QED and our subsequent observations), we do not expect our explanation of gravity or our explanation of snowfall in the Alps to contradict it either explicitly or implicitly. Moreover, we expect the explanation of one event to apply to other relevantly similar events. We don't expect all experiments with water to be explicable under a certain theory, except for

one particular kind of water available from a Fountain of Youth somewhere in South America. That is, we expect explanatory theories to be both time-independent and location-independent.[13] So suppose there is some class of theories to explain why my sensations are correlated in the way they are, *ex hypothesi*, with my neural activity. Far from resigning ourselves to scientific impotence on the grounds that we've really only got one experimental situation before us, we should be happy to conclude *at the least* that some of these theories are going to apply similarly well to other similar neural systems. (In the following section, we discuss what might make systems *relevantly* similar.) Just as in the case of any scientific endeavour, we should expect that the correct explanation of a *single* observed phenomenon—whatever that explanation might be—does its job for more than just the single observed phenomenon at hand.[14] Indeed, that an explanation does have a wider applicability than some special single case is one of the criteria for deciding it is a good explanation.

Moreover, let's consider an argument similar to this one turned on its head. Suppose there is some class of theories to account for all there is to know about all the *other* brains and minds—if the latter exist—in the entire cosmos except mine. (Hopefully this is a fair supposition, although some might of course suggest that the class of theories accounting for all the other brains and minds is empty or that all the theories account for minds entirely separately and independently from brains.) Insofar as *my* brain is similar to the other brains, we should expect that I also have (or lack) a mind in accordance with the same explanatory theory which accounts for all the features of all the other organisms with brains. And since the other organisms' lacking a mind would imply *my* lacking a mind, and since I know that I have a mind, we can immediately infer that the other creatures possessed of similar brains also have minds.

Notice that all these characteristics of explanatory extrapolation in scientific investigation are independent of whether, for instance, in the

---

[13] Of course we needn't exclude something like the possibility that the gravitational constant might change over time or the reality that people can run a marathon in air in a couple of hours but they can't do the same under water. Well formed theories take these factors into account as independent variables.

[14] Equivalently, when extrapolating accounts of unobserved systems on the basis of observed systems (or the single observed system, in our case), the scientist expects continuity over at least *some* surface in the variable space of the type of system under study. That is, the expectation is that a single theory will do for all systems more or less similar to the one under study.

end our preferred explanation of the correlations in the first person case (or the third, as in the second argument) happens to involve resorting to Cartesian dualism or whatever. These observations do not presuppose anything about materialism: if we are happy to incorporate Cartesian dualism or any other *ad hoc* premises into our understanding of the relationship between neural activity and sensation in the first person (or the third), then we ought to be perfectly happy to extrapolate it to our understanding of that relationship in the third person (or the first). As long as we're willing to grant that there *is* some theory to account for the correlations—whether or not we know that theory—then we ought to be happy with the extrapolation.

3.5.2 On a Good Argument from Science

So, what can this tell us about the third person problem, the notion that no phenomenological description of a sensation can be derived from an objective description of a third person system claiming to be having that sensation? Given that we cannot make the details of the complete neurophysiological behaviour of another system available to the right level of our own systems, what extrapolations can we make from the correlations between our own neural activity and our own sensations to the claims of a third person? (I deliberately use the noun 'claim' here with the heterophenomenological[15] detachment popularised in Dan Dennett's work.) Just as in the first person case, we shouldn't look for any kind of deductive logical derivability here, but if the above account of extrapolation is correct, at least to a first level of approximation we should be able to correlate our own experiences with those of a third person to an extent determined by the degree of relevant similarity between our own and that third person's neural activity. I have argued above, *contra* most commentators, that *some* kind of extrapolation from our single observed first person case to the case of third persons is perfectly legitimate in the context of scientific inquiry. The problem now is to investigate what makes cases *relevantly* similar.

First, I suggest we are justified in limiting the scope of possible relevantly similar characteristics to the physical observables (in the quantum mechanics sense—see the chapter following) at the lowest level

---

[15] At ten syllables, 'heterophenomenological' wins first place on my list of longest useful terms in philosophy and cognitive science.

and whatever other organisational features may derive from the physical observables at higher levels. While this is a highly significant move, it may be be undertaken for no more obtuse a reason than that this is the scope of relevant similarity for all other current scientific accounts—with which we may expect a good explanation of neuron-sensation correlations to cohere. (It is here that our hitherto accommodating attitude toward alternative, nonmaterialist, accounts of sensation begins to show signs of fatigue.) Given the types of explanations with which we've been happy in science so far, we should be no more concerned to incorporate nonmaterial factors into our understanding of sensation than we are to incorporate them into our understanding of the electroweak force.

Second, I suggest that given our apparent utter insensitivity to the activity of single neurons, relevant similarity is probably to be found at no lower level than that of the organisational structures embodied by neurons (or whatever information transforming elements happen to make up the system in question). While we will have much more to say on this topic in the coming chapters on self models (see also the chapters on levels of description in cognitive systems), for the moment suffice to say that it would seem peculiar in the context of other scientific theories to need recourse to information about specific individual neural activity in order to explain such an apparently high level feature as sensation. Although it is necessary to understand what enables fusion between two or three atomic nuclei in order to understand an account of supernovae, for instance, it is only necessary to appeal to much higher level features of *structures* made partially of nuclei undergoing fusion in order to give a comprehensive account of what can make a star explode. Likewise, it is reasonable to think that whilst we will need to understand what makes individual neurons fire and interact the way they do, a complete account of what makes someone specifically capable of sensation probably (only?) can be given at a much higher level of description.[16]

Given these two restrictions on what makes systems relevantly similar, and keeping in mind the initial observations about extrapolation and continuity, a sharper outline of a response to the third person problem comes into focus. Above we wanted to say that we should be able to correlate our own experiences with those of a third person to an extent

---

[16] And it's a good thing, too, lest we find our sensations constantly disturbed by the effects of cosmic rays passing through our brains and the ongoing death of our nerve cells.

determined by the degree of relevant similarity between our own and that third person's neural activity. Now this can be rephrased as the response that the claims of another system to be having sensation can be evaluated by comparing the organisational structures physically embodied by the other system to those embodied by our own. Where those organisational structures are quite similar (as in the case of humans), we may be justified in understanding the third person's sensations to be quite similar to our own; where they are quite different but not altogether alien (as in the case of bats), we may be justified in understanding the third person as probably *having* sensations but of some quality of which we cannot be certain; and where the organisational structures are entirely alien or—perhaps more importantly—vastly simpler (as in the case of my notebook computer running its standard operating system), we may in the first instance be unable to judge, or in the second we may be justified in understanding the third "person" to have no sensations whatsoever. (A better specified account of some particular kinds of organisational structures which may support finer distinctions than these is on offer in the coming chapters on self models.) That this is a sort of ultimate *hermeneutische zirkel* I believe is inescapable: we can never see or even evaluate the sensations of other systems save through our own eyes and the backdrop of our own complete systems. Yet this is no criticism of the approach I've outlined, at least not any more than it is a criticism of, for instance, scientific inquiry in general that we construct explanatory theories always in a form consistent with the logic we ourselves find in our own minds (perhaps after some reflection) compellingly rational.

3.6 The Rose Named

To sum up, we saw first that much of the confusion over what is "missed out" of physical accounts of objects of sensation or even physical accounts of neural systems experiencing sensations comes down to a confusion over the levels at which physical information may be processed. We observed how appropriate use of physical information about objects of sensation may lead us to empirical discoveries about how those objects smell, look, or whatever. Use of the identical information in other ways may not lead to any discoveries at all about their rôle in subjective experience. That these two observations can each reflect true

characteristics of the world appears to be an accidental property of human neurophysiology.

In terms of what we have dubbed the first person problem and the third person problem, we've returned to the kind of position favoured at the end of the original discussion of Nagel's "what is it like..." argument. The fruitful approach is one in which objective properties of systems are open to investigation and the correlations between these objective properties and subjective sensation are, in the first person, matters for empirical study. In the third person, the temptation to scepticism because of the impossibility of experiencing someone else's sensation is pacified and the matter is again open for empirical study and legitimate extrapolation from the first person case.

It is worth noticing that in terms of empirical investigation of the relationship between particular organisational structures and sensation, instead of asking *why* particular neural activity feels the way it does, we should just accept that it *does* and get on with the task of exploring *how* this or that perception or sensation is enabled by such and such an underlying organisational pattern. As suggested in the second chapter, we should be concerned with giving neurological descriptions of the structures which enable particular cognitive tasks to be performed but not of *why* those cognitive tasks feel the way they do when performed with those structures. Although it is an elementary point of philosophy of science, it is rarely acknowledged in this context that every scientific inquiry must at some point pass from trying to explain the *why* of observed phenomena to explaining the *how* of the phenomena at a lower level of description. That is, we may give a reductive account of observed phenomena in terms of (perhaps unobserved) lower level features, but at some stage those lower level features themselves cannot be accounted for by appeal to yet lower level features and must simply be described as matters of fact.

Indeed, the reason the reductive accounts of science work as they do in the first place (and not in some other way) is not given a *why* account but is merely described. For instance, we can answer the why of many features of chemical reactions by appeal to characteristics of atomic structure, and we can answer the why of many features of atomic structure by appeal to the electroweak force, and perhaps some day we will be able to answer the why of many of the features of the electroweak force by appeal

to some lower level characteristic of reality which might unify the electroweak force with the strong nuclear force and with gravity. But first, at no stage of the game does science offer an account of why it is that the electroweak force explains in the way it does (instead of in some other way) features of atomic structure—it only describes the way in which it in *fact* does—and second, at some point (perhaps even in the unlikely event it should turn out that a Grand Unification Theory or a Theory of Everything is a necessary truth of logic) science becomes simply a *description* of the way the world works.

Science may offer a sort of promissory note, which says in effect that the necessity of a given relationship or explanation between levels will become clear as soon as other features at a lower level are elucidated fully, but at some point that promise becomes as empty as that on the front of a banknote: "The Royal Bank of Scotland plc Promise to Pay the Bearer on Demand Twenty Pounds Sterling at Their Head Office Here in Edinburgh". Although an entire economy is built upon sterling, just as an entire physics is built upon the electroweak and other elementary forces, the lowest level banknote cannot be redeemed at a bank for anything except another banknote.

I suggest than in questions of mind, we have already reached that point of "mere" description for the simple reason that at the macroscopic levels of description usually appropriate to cognitive science, there is much less ground to cover. That is, we have already reached the point where we must simply acknowledge that particular organisational structures enable conscious perception or sensation and that this just *does* feel some particular way. The project then is to describe what structures do enable perception or sensation and even how activity in those structures correlates with particular perceptions or sensations. But asking *why* it feels the way it does, or *why* it feels any way at all, is on this account akin to asking why there is something which leads to the appearance of the strong nuclear force (or something which leads to the something which leads to the appearance of the strong nuclear force...) and is no more liable to be answered than the question "why is there something rather than nothing?"

In this same scientific vein, we make a brief side trip in the next chapter into the realm of quantum physics before proceeding with an exploration of the self model approach to understanding the subjective

aspect of experience. Our foray into physics is motivated by the need to guarantee that it is not the haven for old-fashioned mind body dualism which some fancy it to be. With this established, we may then continue with the more immediate matter at hand: the matter underlying the self.

4

Materialism and the "Problem" of Quantum Measurement

We have so far made a case for a brand of dual aspect monism, and shortly we will be prepared to advance to an exploration of a particular way of seating the subjective aspect of sensation within the objective context of a material substrate. But at the same time as we embrace the framework of modern physics as the foundation of our monism, we must guard against a popular flanking attack on this foundation which arises from within the very same physics. In particular, we must be sure that there is not yet some room for the old Cartesian dualist to slip in with an argument from the state vector reduction of quantum mechanics.

The following is reprinted, with minor changes, from Mulhauser (1995 in press).[17] The material is occasionally technical, but in the interest of thoroughness I have elected to preserve its original form as much as possible. For the sake of the general philosophical payoff and the value of the points here in ensuring that we are on the right track in keeping our account of the subjective experience of sensation within the realm of the material, I hope it is not unreasonable to ask for a little patience with subject matter largely couched in terms of quantum mechanics.

For nearly six decades, the conscious observer has played a central and essential rôle in quantum measurement theory. We here outline some difficulties which the traditional account of measurement presents for material theories of mind before introducing a new development which promises to exorcise the ghost of consciousness from physics and relieve the cognitive scientist of the burden of explaining why certain material structures reduce wavefunctions by virtue of being conscious while others do not. The interactive decoherence of complex quantum systems reveals that the oddities and complexities of linear superposition

---

[17] Please see the Appendix for reprint information.

and state vector reduction are irrelevant to computational aspects of the philosophy of mind and that many conclusions in related fields are ill founded.

4.1 Quantum Measurement—The Ghost in the Mechanics

Consider how different life would be if we found ourselves in a world where macroscopic objects like bats, cats, lumps of wax, and even people evolved in time the way sub-microscopic objects like electrons and pi-mesons do under the Schrödinger equation of quantum mechanics. My friend might admonish me, "Hey, I saw one of your state vector components out with that MacLean woman last Saturday—I thought you had better eigenvalues than that!" I might justifiably retort that I was *also* in my flat doing work, and he should localise his wavepacket elsewhere and stop interfering with my superpositions.

The standard account of why we never see objects in states of linear superposition is that the very act of observing a quantum system precipitates a discontinuous jump in the system's state from what might have been a superposition into a single determinate state. In the Hilbert space framework of quantum mechanics, wavefunctions are represented as vectors, and maximal quantum observables correspond to operators. For each of these operators, there is an associated *basis*, a set of orthonormal vectors which spans Hilbert space and represents the eigenvalues of that operator. For our purposes, these eigenvectors can be thought of simply as the real states in which it is possible for an observed system to exist. According to the projection postulate, originally due to von Neumann (1932), when a quantum system is observed the system's wavefunction, its state vector in Hilbert space, is projected discontinuously into an eigenstate of the appropriate observable. The probability of the system's being found in a state corresponding to any given basis vector is simply the square modulus of that vector's coefficient when the state vector is expressed as a linear combination of the basis vectors. The set of probabilities corresponding to the eigenvectors when a given operator is applied to a wavefunction is called that state vector's *reduced density matrix*. The process of state vector reduction when a quantum mechanical system is observed—"collapsing the wavepacket"—has excited the attention of philosophers both because of the indeterminacy the reduced

density matrix brings to physics and because of the high stature it is understood to give to the *consciousness* of the observer.

Under the projection postulate, it is irrelevant to the statistical predictions of quantum theory at what point state vector reduction is taken to have occurred—as long as it is some time before the outcome of a measurement enters the *conscious* mind of an observer. That state vector reduction could take place *after* a quantum system interacts with a macroscopic measuring apparatus but *before* a conscious observer has noted the state of the apparatus is the basis of the well-worn thought experiment about Schrödinger's cat. The example includes some device meant to poison a cat (who is taken, perhaps wrongly, not to be conscious[18]) if and only if a detector measures a certain event in a quantum system such as the decay of a nucleus in a radioactive sample. According to the laws of unitary evolution (i.e., evolution in accordance with the Schrödinger equation), a system like this which is appropriately shielded from the environment—more on this later—must evolve into a superposed state representing *both* the case where the atom decays and the cat is poisoned *and* the case where the atom does not decay and the cat lives. It is supposedly only the act of *observing* the system—opening up the box and peering inside, if you will—which brings it about that its wavefunction description reduces to a single eigenstate in which the cat is either determinately alive or determinately dead. As someone has said of the latter possibility, "curiosity killed the cat".

Because interaction with a conscious mind *bounds* the time by which state vector reduction must occur, and because physicists have understood to be unverifiable any prediction that it occurs earlier, some physicists (perhaps Wigner 1962, 1967 most famously) and many philosophers have taken consciousness *itself* to be the mechanism which brings about wavepacket collapse.

Even in Everett's (1957) many worlds interpretation of quantum mechanics, in which state vectors are never reduced, consciousness nonetheless plays a central rôle. Under his account, the consciousness of observers remains responsible for the perspectival nature of experience—the fact that observers only ever experience one of the many components of the superposition of states through which the cosmos is continuously evolving.

---

[18] As the "Wigner's friend" thought experiment shows, this is not crucial.

However we interpret quantum measurement along the traditional lines, we seem faced with an unexplained consciousness phenomenon which somehow makes everything go. Next I outline some of the problems this ghost in the mechanism creates for materialist accounts of cognition.

4.2 Problems for the Materialist

By the term 'materialist', I mean to include all monists who hold that only physical things exist, that there is no separate realm of mind things with positive ontological status, that the world is not instead purely ideal. I mean also to include dual aspect monists, who maintain that there are matters of fact about what it is like to *be* a given material thing which may not be expressible purely in terms of the objective physical properties of that thing.

Regardless of the particular brand of materialism we are concerned to defend, maintaining that consciousness is a physical phenomenon while allowing that it plays the unique rôle in quantum measurement theory it has hitherto been accorded means giving an account of how it is that conscious material arrangements reduce state vectors while other, perhaps equally complex but nonconscious ones, do not. For instance, a materialist who is a functionalist must explain what particular types of information processing arrangements are capable, all by themselves, of reducing state vectors. (This might lead to something as peculiar as: applying function f^c to datum **x** brings it about that **x** has become conscious—and the state vector thereby has been reduced of the entire composite system consisting of both that of which **x** is a measurement and whatever is doing the calculating—whereas applying any function $f^1 \ldots f^n$ does not.) More to the point, we must answer the question of why some physical systems are, by virtue of the functional arrangements they embody or whatever, prohibited from existing in states of linear superposition while other similar ones apparently are not. But the problem is worse.

Indeed, if the source of consciousness is to be found in functional arrangement, quantum measurement theory implies that we should be able to pin down the exact spatio-temporal location in an information manipulating process where a given piece of data becomes conscious. The

projection postulate does not require that state vector reduction take place *at* the terminus of what has come to be called the *von Neumann chain*, the chain of interactions from quantum system to conscious mind which constitutes an observation. But it does require that there *is* a terminus, such that if state vector reduction takes place *after* that point, then an experiment could be devised to show it. If consciousness can be described in functional terms, then so must be the location of this terminus.

Aside from the bizarreness of effecting state vector reduction of quantum systems by applying functions to data about them, pinning down an exact location where a piece of data becomes conscious should be unacceptable to any materialist who wishes to describe consciousness in terms of *processes* which are not necessarily functions.[19] In this case, there might not be any well defined time at which a piece of data enters conscious awareness.[20] But we are then left with a clumsy notion difficult to reconcile with the mathematical elegance of the rest of quantum theory: an ill defined terminus to the von Neumann chain itself. Moreover, with an ill defined terminus to the chain, it is awkward to accommodate the fact that we are still guaranteed *some* time such that it could be experimentally verified if state vector reduction occurred after it but not before it.

A similar line of thought leads to the unappealing conclusion that consciousness cannot be a vague phenomenon: it must be an altogether all or nothing affair. This is because while the *predictions* of state vector reduction are probabilistic, that it occurs is not. Either interaction with a given physical system forces state vector reduction, or it does not. There can be no fuzzy area in between. Indeed, we could imagine a sort of "consciousness detector" which exploits the familiar behaviour of the double slit experiment. Given a "sufficiently shielded" system akin to Schrödinger's cat arrangement, we might fit an electron measuring apparatus with a (nonconscious) device to convert information about electrons *before* they've passed through the slits into an appropriate form

---

[19] This might be the spread of an activation pattern across a network, for instance. While all (recursive) functions can be thought of as algorithms, not all algorithms are mathematical functions. Functionalists are typically concerned with the broader class of all algorithmic processes. Fortunately, however, something like the more descriptive but awkward "algorithmism" has never entered use.

[20] As an aside, Lockwood's 1989 relativistic argument for a precise physical location of mental "events" rests on the assumption that such events have a precise location in time—an assumption which is untenable on any sort of connectionist or distributed view.

and pass it on to whatever possibly conscious system we're wishing to analyse. We may then simply run the electron gun for awhile, and when we examine the photographic plate, we'll find an interference pattern if and only if the subject of the experiment *did not* consciously process information about the electrons. If the pattern corresponds to that predicted by classical mechanics, then it was because the state vector descriptions of the electrons were reduced as a result of information about them becoming conscious.

Finally, accepting that state vector reduction occurs as a result of interaction with any and only those material arrangements with some special material property that makes them conscious even has curious implications for the way we think about the evolution of conscious life. If conscious life was not present when the cosmos began, then the universe could only have evolved (in the mathematical sense) in a state of quantum linear superposition until the first conscious organism evolved (in a biological sense) and observed it, thereby collapsing the wavefunction of the entire cosmos and making determinate that single path of history which made the organism's own existence possible! We might of course posit a (material?) divine being who frequently observed the cosmos and prevented its ever evolving into a superposed state. Since the phenomenon whereby frequent observations of a quantum system keep it from evolving into a superposed state is often called the "watchdog effect", we might name the divine observation hypothesis the "watchgod effect". But in any case, without such a "watchgod effect", it would appear that the first conscious organism was its own efficient cause.

Fortunately, all these strange problems with including consciousness in quantum measurement theory apparently never need arise. While it is always dangerous to speculate on anything's being "an answer" in physics, it appears that the quest has ended for a theory of quantum measurement which discharges consciousness from its central rôle. The best thing about the new view of quantum measurement is that it requires no new premises: it falls out of a careful reexamination of the problem and numerical analysis of the evolution of complex systems described under *existing theory*.

4.3 Interactive Decoherence—Ghostbusting

The current description of interactive decoherence[21] was originally motivated by quantum cosmology and both benefits and is benefitted by research in the physics of information. Quantum cosmology (see, for instance, Coleman, et al 1991) seeks to understand the entire cosmos as a quantum system. This approach can accommodate neither an arbitrary Copenhagen-style distinction between microscopic and macroscopic worlds nor an unexplained consciousness phenomenon driving state vector reduction. The quantum cosmologist must ultimately be able to derive a description of a quasi-classical world from the laws of quantum mechanics. From a quantum description of the world, we must be able to predict the existence of "correlations" between macroscopic coordinates and momenta which approximately obey the classical laws of motion, and we must be able to account for the fact that interference effects between different classical states are never observed. (Paz and Sinha 1992) The relevant aspect of information theory is the growing conviction that information cannot be abstracted away from a physical substrate (Landauer 1991) and how that fact bears on what can be said about natural laws, observers, and the interactions between subsystems of the cosmos.

The most important step in the development of decoherence theory was the "re-realisation" that no system but the entire cosmos is closed, or perfectly isolated, and that the environment will thus always contain some amount of information about the state of a system. The Schrödinger equation is meant to apply *just* to closed (or very nearly closed) systems, and for the sake of computational simplicity absurd degrees of isolation are often tolerated in examples of the Schrödinger equation's application. (This is the point of the extremely well-shielded box in Schrödinger's cat example: no information about the coherent superposed state of the system must exist in any external system, for then observation of this external system would collapse the wavefunction of the entire composite system.) But numerical analysis of systems which preserve some of those complications abstracted away in the idealised example systems—

---

[21] In the physics literature, this phenomenon is consistently referred to as "spontaneous decoherence". However, as will become clear, the phenomenon is not spontaneous in the strict sense and occurs always as a result of information-carrying interactions between subsystems. Thus, with apologies to the physics community, I have opted to use this more accurate term throughout.

essentially much greater internal and external degrees of freedom—reveals that correlations between the state of a quantum system and its environment or even correlations within itself are sufficient to break the coherence of what might otherwise be an incredibly complex wavefunction.

These correlations are understood as records, or *information*, about the system, information which Wojciech Zurek (1991), a leading researcher in interactive decoherence at the Santa Fe Institute, insists is entirely independent of the presence of any conscious observer. The buildup of nonseparable correlations between the system and its environment (which could be little more than cosmic background radiation) causes a very rapid decrease in the possible superpositions of the system *which can be distinguished through their effect on the environment*. As Paz, et al (1993) put it, "this results in a negative selection which leads to the emergence of a preferred set of states...which remain least affected by the 'openness' of the system in question." (p. 488) It is these preferred states, sometimes called the "pointer basis" (a term coined by Zurek, alluding to the pointer of a garden variety measurement apparatus), which, conveniently and unsurprisingly, correspond closely to those of the observables we encounter in the quasi-classical world.[22] (Albrecht 1992b; Paz, et al 1993) The cosmos is watching! While the dynamics of the system determine the "options" for a system's evolution, it is the correlations between the system and its environment—rather than the intervention of any conscious observer—which determine the probability of the system's being in a given state.

It is important to stress that while analysis of interactive decoherence provides the reduced density matrix, or set of probabilities for each of the possible states "allowed through" the nonconscious environmental record-keeping, it is not, as one researcher has called it, a mere "calculational tool" (Kiefer 1991, p. 379) with which we duplicate the predictions of consciousness-driven wavepacket collapse while never essentially erasing consciousness from the picture. In effect, decoherence *supersedes* the wavepacket collapse of traditional quantum measurement theory by offering an alternative account of what is mathematically the

---

[22] Note, incidentally, that this doesn't imply *all* large systems decohere: as Paz, et al (1993) and Zurek (1991) point out, even a very massive—on the order of one tonne—cryogenic Weber bar, by virtue of its extremely low temperature, must be treated as a coherent quantum harmonic oscillator.

same process, free of the superfluous and unexplained consciousness factor. Indeed, the equivalence of results provided by the two mechanisms leads some researchers to apply the older term explicitly in referring to the replacement process. (Albrecht 1992a; Paz, et al 1993) To apply the point to Schrödinger's thought experiment, decoherence tells us that the cat is already either alive or dead long before anyone opens the box—with a probability given by the appropriate reduced density matrix—but *not* as a result of a von Neumann chain-style interaction with consciousness at the terminus.

Finally, in the interest of thoroughness, I should mention that while the description I have given of decoherence is based purely on existing theory, there is another formalism known as the "consistent histories" approach which does rely upon a "decoherence functional" (Gell-Mann and Hartle 1990) which has not yet been fully defined. It is related to the sum over histories formulation of quantum mechanics and is used to determine whether one can attribute well-defined probabilities to different possible histories of a given system. (When this is possible, the histories are called "consistent", or "decohering".) However, this second approach in its present form allows through as "consistent" sets of highly non-classical histories. For this reason, the environment-induced superselection I have described is preferable. (See Paz and Zurek 1992 for one comparison of the two formalisms.)

4.4 Quantum Mechanics is Irrelevant

We can see from this description of interactive decoherence that the consciousness of an observer is no longer essential to the theory of quantum measurement. As Zurek puts it,

> "Conscious observers have lost their monopoly on acquiring and storing information. The environment can also monitor a system, and...such monitoring causes decoherence, which allows the familiar approximation known as classical objective reality—a perception of a selected subset of all conceivable quantum states evolving in a largely predictable manner—to emerge from the quantum substrate." (Zurek 1991, p. 44)

As it stands, even in the absence of a conscious observer, the wavefunction of any quantum system with sufficient complexity and energy will decohere. Thus it seems that apart, perhaps, from theory concerning very low energy computation, *quantum mechanics is utterly irrelevant to computational aspects of the philosophy of mind.* None of the problems I outlined for materialism in general, functionalism in particular, or even the origins of conscious life arise under this new picture of quantum measurement.

Likewise, many interesting results in the philosophy of mind and related fields which have derived from the assumption that macroscopic objects can exist in superposition until they are observed have lost their theoretical underpinnings. For instance, Deutsch's (1985b) "universal quantum computer", whose capabilities are a superset of those of the familiar Universal Turing Machine or Bernoulli-Turing Machine, seems destined to exist only in the world of theory. The eventual application of other research in quantum computing (for instance, Margolus 1986, 1990) inspired by Deutsch or Feynman's (1986) efforts is unclear; what is clear is that any quantum computer of even rudimentary complexity must operate at extremely low temperatures in order to preserve the coherent wavefunction description on which such devices rely for their special properties. (Indeed, the information processing nature of such devices might, in itself, create such internal correlations that coherent unitary evolution cannot be sustained.) Because the operating temperature of the human brain is many orders of magnitude higher than what is required to sustain prolonged unitary evolution these special properties of quantum computers are almost certainly irrelevant to brain research. Unfortunately, it seems also that in light of interactive decoherence, Deutsch's (1985a) description of an experimental test of Everett's interpretation (a suggestion contradicting the conventional wisdom that it is indistinguishable from rival interpretations) using nonconscious automata is also unworkable. This should not be too surprising, however, since Everett's theory stipulated that state vector reduction never actually took place. While it is certainly no trivial project, we might anticipate that some or all elements of Everett's view will soon be proven inconsistent with decoherence theory.[23]

---

[23] Since this paper went to press, it has occurred to me that Everett's theory mightn't be inconsistent with interactive decoherence after all: the cosmos might simply "auto-select"

Albert's (1983, 1987, 1990) work showing that a specifically nonconscious automaton could make privileged predictions about itself by measuring quantum observables which for any external observer would be incompatible appears similarly incompatible with interactive decoherence. Although arguments from dual aspect monism indicate a necessary subjectivity to the point of view of an observer (see Chapter 2 and Mulhauser 1993a), and Mackay (1971, 1980) has argued for an observer's "logical relativity", Albert's arguments for subjectivity fail because they require complex automata fitted with quantum mechanical measuring devices to themselves exist in states of linear superposition.[24]

Finally, the new view of quantum measurement does not mix well with the mind-brain interaction theories of Sir John Eccles, Nobel prizewinning neuroscientist. Eccles, a self-avowed dualist with respect to the mind body problem, has described a scheme (1986, 1990; see also Popper and Eccles 1977) in which a nonphysical consciousness collapses the state vector descriptions of the pre-synaptic vesicular grids which release neurotransmitters at neural junctions. He proposes that states of columnar bundles in the cerebral cortex thus become correlated with the causally prior mental "psychons" with which they are paired. But not only is consciousness itself superfluous in decoherence theory, the high operating temperature of the human brain again guarantees decoherence of the wavefunctions of these structures as a result of internal and external correlations, independently of any mysterious causally prior mind entity.

In addition to neutralising all these interesting results which come from allowing nonconscious macroscopic objects to exist in superposed states until they are observed, interactive decoherence appears also to have solved the preferred basis problem. This is the question of why Nature has chosen for macroscopic objects a set of basis vectors which correspond to the eigenstates of macroscopic observables. (Why not a basis corresponding to some other set of operators, such that the eigenstates we observe are actually superpositions of the eigenstates of the macroscopic observables? Out of the infinity of ways of decomposing state vectors, what makes the basis corresponding to the set of macroscopic observables so special?) The emergence of a preferred basis simply as that basis which

---

its states every so often in a way so as to reduce periodically its infinite class of superposed states predicted by Everett to a smaller class corresponding to the preferred basis states.

[24] The failure of Deutsch's and Albert's work as physical possibilities also casts some doubt on the practical feasibility of quantum cryptography.

is most immune to the openness of macroscopic systems is at the heart of decoherence theory.

Thus Lockwood's (1989) approach to the preferred basis problem through an unexplained consciousness phenomenon in a reincarnated relative state view is as unnecessary as it is implausible. Interactive decoherence suggests a similarly dim view of Deutsch's (1985a) interesting but apparently only partially successful (Foster and Brown 1988) attempt to solve the preferred basis problem. Other approaches either to removing consciousness from quantum measurement altogether or to solving the preferred basis problem are now also unnecessary. These include Davies's (1981) and Penrose's (1985, 1986, 1989) quantum gravity state vector reduction and Nicholas Maxwell's (1988) propensition theory positing state vector reduction in the wake of sufficiently energetic inelastic collisions between particles.

Overall, the mechanism of interactive decoherence appears to solve a host of problems without creating very many new ones. But the question lingers: is this *really* the way it happens (or at least a reasonable approximation), or is it just a parallel account of the observed phenomena which offers no particular verifiable advantage over the standard consciousness-driven wavepacket collapse? Without entering a prolonged discussion of the philosophy of physics, there are a few illuminating things which we can say about this.

Insofar as both decoherence theory and the standard view yield the same reduced density matrix for the quantum systems so far studied, the huge body of positive experimental evidence for the accuracy of quantum mechanics as a predictive theory tends to confirm both views equally well. Yet, the mechanisms which *precipitate* interactive decoherence come for free as consequences of other elements of existing theory. The same cannot be said for the standard view, which relies upon the superfluous phenomenon of consciousness to terminate the von Neumann chain. In that sense, interactive decoherence is a more parsimonious theory. For that reason alone, independently of possible experimental verification, the standard view may eventually be replaced as interactive decoherence theory becomes more widely understood.

However, at least in theory, it *is* possible to distinguish experimentally between the two accounts. I have said that they give identical predictions for all quantum systems studied so far, but so far not

all imaginable quantum systems have been studied. Specifically, the two accounts predict different outcomes for experiments with the fanciful "consciousness detector" I described above. If such a system—consisting of a standard electron gun and diffraction grating setup, together with an "observer"—could be shielded from the environment, decoherence theory predicts that the consciousness detector simply would not work in the way I outlined under the standard view: the correlations between the states of the electrons, the measuring apparatus, and the "observer" (conscious or comatose) would cause decoherence and yield a classical distribution pattern on the photographic plate every time. The problem here, of course, is that such an experiment requires a fantastic degree of isolation far beyond the technological capabilities of today or the foreseeable future. Probably well before such isolation becomes possible (if it ever does), theorists will determine how better to quantify the *amount* of information which must be carried in inter-system correlations to guarantee interactive decoherence. In that case, a similar test could be carried out by replacing the "observer" with any system capable of interacting with the electrons to the required degree. Until this necessary degree of interaction is quantified[25] (or ridiculously thorough isolation becomes a reality), experimental discrimination between the two accounts will remain practically impossible.

Decoherence theory does not answer all the interesting questions about quantum mechanics—such as why linear superposition ever occurs at all or why experimentally verified nonlocality is an apparent feature of reality. It also raises at least one intriguing new question: could the state of the system's environment, considered in all its detail, influence *which* eigenstate a system's state vector jumps to? My own suspicion is that a new non-local but deterministic picture of quantum reality, more satisfying than Bohm's (1952) and incorporating a fuller description of interactive decoherence, may be forthcoming. But for now, the cognitive scientist and philosopher of mind can rest assured that the burden has been lifted for giving an account of material consciousness capable of

---

[25] Early indications are that the time required for decoherence, and perhaps the degree of necessary interaction as well, are very small indeed. Although he doesn't include the technical details, Zurek says that a rough calculation reveals that for a room temperature system with a 1gm solid mass, quantum coherence is destroyed in less than 10^{-23} seconds. (Zurek 1991)

playing the state vector reducing rôle hitherto supposed necessary in explaining the observed quantum mechanical phenomena.

4.5 Interactive Decoherence—An Afterthought (?)

Soon after making available on the International Philosophical Preprint Exchange[26] a preprint of the *Minds and Machines* article from which this chapter is taken, I received some interesting objections from Nobel prizewinning physicist Brian Josephson, of Cavendish Laboratory at the University of Cambridge, on positions I have taken on points which really sit at the very centre of present debates about decoherence. As he suggests, there does often seem to be some sleight of hand at work in the decoherence literature, although he concedes that the merit of my own account of interactive decoherence may be that it goes through the argument sufficiently clearly that perhaps we can see where the sleight of hand occurs! (Josephson 1993b) With that thought in mind, then, let's address his concerns directly and try to be sure that we have discharged any sleight of hand from important rôles in the argument (or any rôles at all!).

The point of the main objection lives in the straightforward question about Schrödinger's cat, "how do we go from the mathematical property of decoherence to the assertion that 'the cat *is* already either alive or dead long before anyone opens the box'?". (Josephson 1993a, quoting me) As he indicates,

> "The nub of the matter is that ordinary physics implies a deterministic correlation between whether the particle decayed and whether the cat is subsequently alive or dead, plus the fact that owing to the linearity of the Schrödinger equation, once a superposition always a superposition. ...Decoherence implies [only] that the two dead/alive components are entangled states [i.e., that the system is in a mixed state—G.R.M.] rather than simple product states." (Josephson 1993b)

---

26 This service, at phil-preprints.l.chiba-u.jp, is generously provided by the Cognitive Science Department of Chiba University, Japan and is mirrored by dozens of gopher and FTP servers across the world.

In an example later, he clarifies with an indication toward what may be the real question, which is whether we can have "continued superposition" when coherence has been lost (Josephson 1993c), and he objects that "the idea that the system is actually *in* one of the... [basis]... states is put in as an *ad hoc* axiom, justified by its consistency"[27] (Josephson 1993d). In other words, decoherence may provide a *basis*, but it doesn't provide any argument why the system must be in one and only one state corresponding to an eigenvector in the basis. Josephson is concerned that the system's actually *being* in one of these basis states is merely added into the account as an unargued afterthought.

The insistence that "once a superposition, always a superposition" (shared, incidentally, with Wigner) is one with which we won't here quarrel, because as we will see momentarily it probably can be sidestepped with the same kind of friendly verificationism to which we appealed with the consciousness detector above. That a system may objectively exist in a superposed state after coherence of the wavefunction has been destroyed is a possibility not even entertained by any of the physicists on whose work this present view is based, but perhaps that is due to "sleight of hand", an assumption that the system is actually *in* one of the provided basis vectors. So, the central question does come down to whether this assumption—that the system is actually in one of the interactively decohered states—really does amount to an afterthought.

Let's consider what might be the experimental difference between the proposition that a decohered system has actually "collapsed" into an eigenstate and the proposition that it exists still in a superposed state, except that the superposition is now (on account of decoherence) a linear combination of the vectors describing only quasi-classical basis states. The first option offers a theoretical account of the system's evolution which proceeds through interaction with an environment and ends with a description of the eigenstates in which it is possible that the system may be found upon observation, together with a prediction of the probabilities of finding the system in any one of the eigenstates. Crucially, the

---

[27] As an aside, Josephson also objects that this consistency is not exact, because we are assuming off diagonal terms of the reduced density matrix are zero, when in fact they are only vanishingly small. A question related to the off diagonal terms in the reduced density matrix was also raised by a woman attending an April 1993 presentation of this material at the University of Edinburgh, but as we didn't follow up the thought at the time, her identity unfortunately remains unknown to me.

probabilities describe the chance that the system will have *already* "collapsed" into one of these states—although, until the observation is made, we remain ignorant of which of the states is real. The second option offers a theoretical account of the system's evolution which proceeds through interaction with an environment and ends with a description of a superposition of eigenstates into one of which the system may be forced by conscious observation, together with a prediction of the probabilities of the system's being forced into any one of the eigenstates. Crucially, the probabilities here describe the chance that the system *will* "collapse" into one of these states—since before the observation is made, the system has not yet collapsed into any of them.

In both cases, of course, the probabilities sum to unity, so the prediction is that the system *will* be found in precisely one of the eigenstates. In thousands of experiments, these predictions have been amazingly accurate: indeed, the precision of these predictions is far higher and far more thoroughly verified than the precision of any other physical theory ever humanly known. But the problem is that *either* account above can get us from start to finish in an experimental observation, and *either* is consistent with the enormous body of experimental evidence for the accuracy of the resulting probabilities. In other words, there doesn't appear to be any *experimentally verifiable* difference between the two accounts. If this is true, has the advocate of interactive decoherence succumbed to the afterthought interpretation and simply opted for the new view over the established one for no sound reason?

We are now in position to answer this question in the negative. If we *begin* from the standpoint of the quantum measurement theory which has been accepted for the last six decades, then "adding in" a proposition that a decohered system is actually *in* an eigenstate before an observation is made does look like it is unfairly putting consciousness on the dole. But if we return to the original projection postulate and the von Neumann chain from above, we may take an alternative view. Recall that consciousness *terminated* the von Neumann chain; that is, the observation was the *latest* time by which the wavepacket could collapse—and recall our note that it is apparently *unverifiable* whether collapse actually takes place sooner. Interactive decoherence now offers us an account—which follows from existing elements of quantum theory—of the actual "collapse", although we still don't *know* the outcome until the

von Neumann chain is terminated. But this is hardly mysterious! We don't know the state of *anything* until we observe it or observe some consequence of it. We now apparently have an account of the emergence of the basis vectors—as Josephson concedes—but it is perhaps confusingly obvious that we can't expect to *know* the actual eigenstate until we observe it. If we accept the above account of interactive decoherence, we are left needing only to answer the question of whether the system is actually *in* an eigenstate before observation. But as we have seen there is no experimentally verifiable difference between the two alternatives, and on this view it is the proponent of accepted quantum measurement theory whose "sleight of hand" is adding in a consciousness phenomenon which has *no explanatory value.* Consciousness is redundant. As far as I can see, on this particular matter of interpretation, the matter—and the mind—is settled.

4.6 Interactive Decoherence—After Afterthought

In fairness to the issue, it is worthwhile to note that still all the questions haven't been answered. In particular, as Josephson has wondered in passing (Josephson 1993d), and as many people may wonder when analysing the account of interactive decoherence here on offer, isn't there some problem in just using the same old reduced density matrix to extract the predictions from decoherence theory? That is, mightn't there be some hidden assumptions in using the matrix which make it altogether unremarkable that quasi-classical basis vectors are automatically accounted for? Is it just a case of the snake biting its tail, without real explanatory value? Unfortunately, this is an issue which I have not yet been able to pursue, and I am unaware of the work of any other theorists who may have answered the question. Hopefully it can be answered by a straightforward reexamination of how we get the reduced density matrix in the first place. As I understand it, there are actually no hidden assumptions in deriving the reduced density matrix in the first place, but whether there is some unexplained element to its connection with decoherence theory remains to be seen. Although everything looks good for the moment, we are still well served to remember the admonition from above that it is dangerous to speculate on anything's being "a solution" in physics!

5

To Be or Not to Be, That is the Data Structure

It is time now to refine our formulation of the kind of dual aspect monism which we dubbed "cybernetic realism" in the first two chapters on objectivity and subjectivity. I would like especially to distance the present position from the naïve notion that there is a mental aspect to every material entity or process. We will not suggest that it feels like something to be any material substance at all (however impoverished that feeling may be, like the photon who perhaps feels either up or down and nothing more); instead, we will examine the idea that it feels like something to be particular kinds of materially instantiated dynamic data structures which we'll call *self models*.[28] (Presently we'll have more to say about what constitutes a data structure.) This is a refinement on the cybernetic realism position, where we indicated that it was like something to be particular kinds of material entities; now we are suggesting that it is like something to be particular kinds of materially instantiated data structures. To put it very crudely, the *I* does not lie in the material entity *per se*, the *I* lies in the self model which may be instantiated by the material entity.

Whether, from a linguistic point of view, this is a legitimate point of departure for trying to get an empirical grasp on the subject of conscious experience is a matter which we shall not debate. I eschew entirely ridiculous arguments of the form, "I have a pain in my tooth; instantiated

---

[28] I am grateful to Sue Blackmore (University of the West of England) for the conversations which first inspired my exploration of self models. The basic idea of locating consciousness in such models originated, in its present form, with her, and appears very sketchily in Blackmore (1993). See especially her Chapters 7, 10, and 11 for applications of the self model idea to explaining problems in psychology at a higher level of description than we address here. The idea has also been taken up by the mutual friend Thomas Metzinger (Zentrum für Philosophie und Grundlagen der Wissenschaft der Justus-Liebig-Universität), who has made some initial attempts at formalising self model descriptions. (Metzinger 1993)

data structures do not have teeth; therefore I am not an instantiated data structure". Someone sympathetic to this argument must work harder than that to convince me that I don't know of *many* examples of instantiated data structures with teeth, including myself. (Of course, I also know of many instantiated data structures without teeth, including some boxers, some old age pensioners, and the list of phone numbers I keep on my computer.) For the most part in this dissertation, we shall be concerned with getting at "what makes a system go" and not with sorting out how to fit our language to that reality or why our pre-theoretic applications of language might be misleading, and we certainly won't engage in prejudging central problems of philosophy with linguistic tricks. Linguistic smoke and mirrors, after all, do not make for good philosophical reflections.

We can hardly pretend that the word 'I' or the word 'self' is sharp and well understood by linguists, philosophers, or psychologists, and we will not here tilt at the windmills which may emerge out of confused and perhaps inconsistent sets of propositions about what makes a self. Instead, I hope the picture which will emerge from this way of viewing the self may improve our understanding of the sense of the word itself. Perhaps under this approach the concept of self will be both better defined and *different* in some key respects to how it might have been seen before. Given more space, we might begin to explore the many implications of locating the self in an instantiated data structure, but for now it will do simply to examine some basic rationale behind it and some of the cognitive neuroscience to which we may appeal to render the idea attractive as more than just a philosophical ruse.

5.1 If Not a Data Structure, Then What?

Having said that, it is useful at least to situate the data structure view of selfhood within a rather meagre context of other possible views of the self which might grow out of the dominant approaches to philosophy of mind. The classic approach, of course, is mind body dualism, and at first blush it offers the clearest idea of what it is to be *me*: the self is simply a non-spatial ethereal spirit "inhabiting" a material body with which it is for now somehow linked. This view packs all the interesting questions about conscious subjects into a nonmaterial spirit beyond the grasp of

objective third person empirical inquiry and, while intuitively appealing, it does less to answer questions about the self than it does to cordon them off into an unbreachable fortress of mystery.[29] Curiously, materialist views often have more in common with this dualist tradition than most hard-headed cognitive scientists like to admit. (And the self model approach is no exception, although we will be more up front about our own "dualist" leanings!) Both so-called "strong AI", for instance, as well as the bulk of functionalist approaches to mind, emphasise the importance of what a material substrate is *doing* over the material substrate itself. In strong AI we are concerned with what algorithm is being instantiated, and in functionalism we are concerned with what functional relationships are being maintained in a complex input/output system. In both cases, the approach amounts to a sort of hardware/software dualism. Within these approaches, questions about the self are rarely broached, but when they are, they ought to be addressed with analytical tools from both the hardware and the software side of things. So far, however, neither camp has mobilised these tools to produce any profound comments on what it means to be a subject of conscious awareness.

Presently we will engage these problems head on. Our approach is not that of either strong AI or functionalism or even of Cartesian dualism, but in making sense of self models there are lessons to be learnt from both the successes and failures of the other approaches. Exploring the self model approach requires bringing the self out of the Cartesian closet and digging through the technical flash of functionalism and related areas of cognitive science, where all too often neurophysiological and psychological data paired with a little computer programming may numb an audience into an overly optimistic impression of the progress being made on really tough philosophical questions. Whatever conception of the self we bring to the discussion must be temporarily set aside so the self model picture can be brought into focus without undue influence from other (underdeveloped) views. In what follows, it may be useful even to pretend for a moment that we have encountered some human who has for all his or her life thought of the self as an instantiated data structure. Where our philosophical intuitions clash with those of this hypothetical

---

[29] While it is hardly necessary to offer references for such a vast literature as that surrounding the dualist approach to the mind body problem, its thorough and systematic destruction of dualism from the inside out merits Smith and Jones (1986) special mention.

individual, what questions would we ask of them to see where they might have gone wrong? How would we convince them that our way of thinking, whatever it might be, is better? What, if anything, is wrong when they say, "I am a data structure"? The Birmingham philosopher and computer scientist Aaron Sloman is fond of saying that people who refuse to accept something like a functionalist approach to cognition are simply in need of long term philosophical therapy. Appropriating his saying, perhaps in the end we will have some opinion on whether it is our hypothetical self model individual or someone with a more standard view who might be in need of such therapy.

In what follows, I have adopted the strategy of first describing in broad strokes what kinds of things we mean by self models and what some of the immediate consequences might be of seating experience in them. We discuss some ramifications of the self model view as if a full understanding of their material instantiation were a *fait accompli.* I ask for a little philosophical indulgence to explore the explanatory power of the idea and some of its consequences before fully fleshing out how we (or a brain) would go about implementing such structures. I hope the sceptic will not abandon the project in the early woolly stages and will reserve judgement at least until some of the more detailed technical explorations to follow!

So, before embarking on an exploration of the particular neural mechanisms which may be involved in the material implementation of a dynamic data structure called a self model, let's take a tour through some of the characteristic properties of such data structures and make some slight refinements to the sense in which we above located the *I* in the self model.

5.2 Self Models Have More Fun

It should be clear from our discussions of the cybernetic realist's brand of dual aspect monism that many questions remain unanswered concerning the relationship between the activity of material structures and the subjective experiences with which such activity may be correlated. One central question, of course, is how *sensation* itself arises: how is it that the subject of conscious experience has any experience at all? What makes the human experience of a flying mallet different than a table's

experience of the same? Moreover, why does a subject experience a sense of colour or of taste or of smell rather than a sense of, say, neural firings in the neocortex? And along the same lines, how can a sensation be continuous when neurons are discrete units? Secondary questions, barely less baffling, concern the process of learning to categorise perceptions in a world free of *a priori* labels, the development of sensori-motor coordination in somatic time, and the rôle of imagination and procedural memory in the experiential landscape of a conscious being.

We may illuminate all these kinds of questions by appeal to the concept of self models which we have just introduced. Centrally, we can interpret sensations (or other conscious experiences) as changes to the self model data structure. That is, to say that I am seeing red is to say that my self model is undergoing a change brought about by neural responses to certain radiation impinging on the retina. Cognition can be understood in the same way, so that to say I have just mentally added five and seven is to say that my self model has undergone a certain transformation brought about by changes in the firing patterns of my neurons. We shall outline in the next section some of the crucial features of self model data structures which distinguish them from other data structures, but for the moment, just to get a handle on what we mean by 'data structure', we may appeal to a digital computer example.[30]

One data structure which is found in one form or another in almost all modern computers is the stack. In the abstract, a stack is a linear structure where pieces of data can be "pushed" one after another for temporary storage; later, they can be "popped" into a central processing unit register according to the LIFO (last in, first out) protocol. The contents of the stack may be changed directly by the programmer by means of instructions which push and pop data; alternatively, they may be changed by the central processing unit itself when, for instance, it encounters a branch or interrupt to a subroutine. In this second case, the processor pushes a note of the next memory location to process for instructions after the subroutine is completed. To return from the subroutine, the CPU

---

[30] It will become progressively clearer that the self model view on offer here is in no way wedded to the classical computing paradigm and is in many ways incompatible with it; despite the use of the 'data structure' moniker, the view does not rely on the data processing or so-called instructionist paradigm, and our data structures needn't have any resemblance at all to the well-defined symbolic data structures of classical computing. (See also Chapter 11.) Nonetheless, the field offers a wealth of examples for thinking about the general concepts to which we appeal.

pops the memory location back off the stack and resumes taking instructions from that point. *As a data structure*, the stack is acted upon by the "external" influence of the CPU. Although the stack is of course a seriously impoverished example of a complex data structure, the example illustrates the sort of thing we mean when we say a dynamic data structure is undergoing a change.

Notice that as a data structure, the stack may be implemented in any number of different ways on any number of different hardware architectures. If we look inside five different makes of computers, we may find the stack implemented in five entirely different physical locations and by altogether dissimilar types of hardware componentry. At a higher level, the organisation of the pointers or handles which are part of the instantiation of the stack—and which themselves might be thought of as an even more impoverished kind of data structure—may also be very different on different machines or under alternative operating systems. Notice also that as a data structure, the stack is ignorant of details of the activities of the CPU (although it is altered by it), and it is ignorant of all the low level aspects of the machine's hardware such as the physical placement of memory chips and the flow of electrons through flip flops. We will soon see in more detail how both these characteristics are shared with self model data structures. Having just appealed to a digital computer example to elucidate the idea of a data structure, however, let's first make some brief notes about consciousness and data structures from the vantage point of what will be an ongoing theme of this material: explanation of observations at various *levels of description.*

5.3 Self Models on Top

One of the most important points to keep in mind with respect to the present kind of analysis of consciousness is that by seeking to understand high level features of consciousness through reductionist *methodologies*, we are not seeking somehow to reduce the features themselves. By locating consciousness in data structures and examining the properties of those data structures, we are just trying to see how the higher level features emerge from lower level interactions and *not* trying to show that the higher level features have no real existence or

importance apart from being a handy way of describing a host of lower level interactions. An example from mathematics is useful here.

Suppose we would like to understand the process of raising a number to the power of a real number exponent. There are many observations we might make about exponentiation and the relationships between different exponentiated bases and about the rôle of exponentiation in other kinds of mathematical processes. But the most straightforward way of understanding the *mechanics* of how to compute the result of raising a base to a real number exponent requires an understanding of logarithms. Understanding the mechanics of computing logarithms (as opposed to understanding how to use them in other processes) again requires understanding the mechanics of multiplication, which in turn requires understanding the mechanics of addition. Yet real number exponentiation is an important abstraction in its own right, and it is a process which does not emerge at all without a sufficiently complex set of (at the lowest level) "mere" additions. The complexity of real number exponentiation is not identical to the complexity of addition; it is an emergent feature—which is "more than the sum of its parts"—of multiple additions put together in the right way.

Something very similar, I would argue, is true of consciousness. While it may be implemented at the lowest level by neural interactions, that does not mean it does not have important features at higher levels of description which really wouldn't make any more sense if expressed at a lower level than, say, a complicated real number exponentiation would make if expressed at the level of repeated additions. Moreover, in the case of something like a self model instantiated by a brain—and this is slightly different to the mathematical example, where often a theorem may be re-cast as an axiom and the original axioms proven again as theorems from the new axioms—there may be features of the dynamics of low level structures which only make sense when they are seen as elements of a higher order structure. This notion invokes no spooky downward causation or some such: it is simply akin to the fact that atoms within a molecule, for instance, may evolve over time through paths which they would never take (or, at least, would be very unlikely to take) if they were not parts of the molecule.[31] The molecule "abstraction", although it

[31] This statement is meant to be taken at face value for the present context and not as a substantive comment on the relationship between physics and chemistry.

ultimately comes down to organisations of atoms, is a useful higher level of description which may be applied in explaining observed behaviour. The same is true of a binary star system: we don't observe lone stars moving in the kinds of orbital patterns through which stars in a binary system move. The two levels of description ("binary star system" and "this star here" together with "that star over there") offer complementary vantage points from which to explain observed astronomical behaviour. Indeed, it is difficult and uneconomical to try to understand everything about the observed stellar behaviour without knowing something about each of at least these two levels of description.

With this *caveat* in mind, we may shortly move on to characterising features of the self model data structure. First, however, we explore briefly two consequences of taking the data structure as a point of departure for understanding conscious experience. These relate to ubiquitous questions about qualia and to the "blindness" of the self model to the substrate (neural or otherwise) which may instantiate it.

5.4 Tiresias was a Self Model

It is significant that in seating experience in the evolution of a data structure, we have linked the phenomenal *quality* of an experience to the capacity of the object of experience to bring about change in the data structure which is the subject. Given the extent of the debate in philosophy of mind over the existence of non-relational qualities of sense experience, or *qualia*, and given the importance of this point for understanding the thrust of cybernetic realism, let's take a closer look at what this means. For the moment, let's assume that any change in the self model as a data structure is functionally relevant to the system concerned.[32] (That is, changes to the data structure are not gratuitous from the vantage point of the system, just as changes in the stack are not gratuitous from the vantage point of the computer.) This view then suggests that if the functional rôle of a given experience were different—i.e., if the object of experience changed the self model data structure in a different way—then the quality of that experience would be different. And

[32] This is not to say that every change in the material instantiation of the data structure is functionally relevant. The point is just that, as we see in the following chapter, the self model typically reflects evolution of functionally important subsystems; thus, it follows that changes in the self model typically indicate changes in such subsystems.

the relationship works the other way as well: if the quality of the experience is different, so too must its rôle in the system be different. (After all, to say the self reacts to something differently is in this context just what we mean when we say the quality of an experience is different.) Since the quality of experience is, on this view, a property of the change in the data structure—and given the restriction that any change to the data structure is functionally relevant—there is no sense in speaking of the quality as something unrelated to the functional rôle of the experience.[33] Thus, on this view, qualia are not non-relational. (Or, equivalently, if we understand qualia to be strictly non-relational, then qualia do not exist.)

We can elucidate the point with a line of thought similar to Dennett's (1988): could we conceive of our pain qualia suddenly changing and being more like our pleasure qualia (perhaps with the two changing places, as in "inverted qualia" examples)? Perhaps we could, but on the present view, this would require that there was also a change in the functional rôle of pain. The reason is simple: regardless of the quality we might like to imagine being associated with pain, if pain still fulfils the same rôle, then we must still find ourselves reacting to it in the same way (disliking it, wanting to avoid it, associating and comparing it in the same way we always have with other sensations, and so forth). Conversely, if we find ourselves reacting to pain in a different way (liking it, seeking it, associating it differently with other sensations or previous pain sensations, etc.), then not only is it no longer pain, but clearly its rôle in the data structure is also different. It does not make sense to speak of the quality of pain turning to the quality of pleasure, if we still don't like it and still respond to it in exactly the same way as ordinary pain! (Consider: "I have agonising chronic pain in my back, but it's alright, since I've had a special operation so that I rather enjoy agonising chronic pain.") This is akin to Dennett's coffee taster who doesn't know whether coffee tastes the same as it used to but he now no longer likes that same taste, or whether the taste of the coffee is actually now different and he doesn't like the new taste. Apparently it is not only true that we could not verify an answer to

---

[33] Of course, a given cognitive system might go on functioning in just the same way, regardless of some tiny difference in the "quality" of two possible stimuli, but if that system were able to *report* this tiny difference, then it would be functioning differently with respect to the two. (It's challenging to search for ways of saying anything meaningful about such differences, or even about a single stimulus, without appealing to some kind of ultimately relational description. I, for one, am at a loss.)

such a question if one were given, but it doesn't even make sense to ask the question. Notwithstanding Puccetti's (1993) criticism in a related context that Dennett (1991) misunderstands verificationism and offers a poor justification for his verificationist tendencies, we can at least be sure that it doesn't make sense to speak of there being any difference between the two possibilities *in terms of the coffee taster's experience.* Is his experience the result of a taste bud change or a preference change? There might be a matter of fact as to whether the coffee taster's taste buds have changed or whether something else in his brain has changed, but if his self model data structure changes the same regardless of whether the taste buds are different or the brain is different, then there is no difference to experiential quality.

Thus, with the cybernetic realist line developed earlier, we are not in the business of trying to attribute some kind of epiphenomenal quality to the experience of systems with brains or other appropriate material structures. Cybernetic realism merely allows us to attribute experience to the kinds of data structures we are presently exploring. It is the position that there *is* a quality of experience to changes in certain data structures, even if that quality can only be described relationally, and even if the "feel" of that quality cannot be logically deduced from objective descriptions of the material substrate instantiating the data structure. Without such a position, we would still be stuck in the first person wondering if other data structures just like our own felt anything at all. With respect to qualia, like Dennett I am *not* suggesting that there are *no* qualities of experience but only that these qualities of conscious experience remain relational. This coheres with Shoemaker's (1975; see also Davis 1982b) early analysis of qualia which suggested that if experience devoid of qualia (zombies again!) were empirically indistinguishable from experience with qualia, then qualia would have no causal rôle, and we could not be introspectively aware of them. Along the lines of Carleton's (1983) analysis, this view is probably essentially correct, despite Block's objection (1980; see also Shoemaker's 1981 reply) that qualia could play a causal rôle which was independent of the functional description of a given subject. (See Tye 1993 for an interesting but probably inadequate comment on evidence from neurophysiology and "absent qualia".)

On a similar note, in seating experience in the evolution of an instantiated data structure, we have eliminated any relationship between

the quality of that model's experience and qualities of the material substrate instantiating it—except for the extent to which properties of that material substrate affect the evolution of the data structure. Just as the stack as a data structure is blind to the peculiarities of the hardware implementing it, so too is the self model blind to the properties of its material substrate. If something makes no difference to the evolution of the self model, then it makes no difference to the experience of the self model. If the same data structure could be instantiated by two systems, then neural wetware would feel just the same as silicon hardware! (see Cuda 1985)

Further observations about the consequences of the self model approach, including important conclusions for the so-called grain problem (how is continuous conscious experience implemented by discrete neurons?) and a variety of the frame problem (how do we know what pieces of data are relevant to completing a task at hand?), must await a more detailed explanation of the actual mechanisms we would expect to see in self model data structures. Moving in that direction, in the next two chapters we examine in more detail some characteristics of the data structure itself, see how it may be distinguished from other data structures which may be implemented by the same underlying material structures, and compare the self model view to functionalism.

6

How to Find a Self Model in a Crowd

Perhaps the most noteworthy property of the self model data structure is that it is a functional representation of the entire system of which it is itself a part. Exploring this idea will be the central task of the next two chapters. We will also explore the rôle of compression in such a data structure and the notion of a "virtual window" which comes from it. We will see that a self model couples the loss of information about low level process with the emergence of new information about higher level processes, and we will note the importance of the polymodal associations this makes possible and give a sneak preview of the architectural reentrance which implements them. The place of environmental feedback in such associations and the significance of the self model's nearly arbitrary combinatorial capacity will become clear. Finally, we prepare the way for the next chapter's exploration of the relationship of the self model view to the broad class of theories of functionalism and discuss what we are *not* saying about the merits of such theories. Incidentally, we omit here a number of basic properties which are fairly obvious characteristics of the kinds of systems we are concerned with. For instance, the material instantiation of a self model must have a degree of plasticity, since otherwise it could not "remember" or change with respect to its environment or its own internal factors. I trust it will not be a criticism of the view on offer that such basics are missed out entirely for the sake of devoting our attention to more interesting material!

6.1 The Self's Self Reference

The basic property of the self model, from which the name derives, is its functional representation of the entire system of which it is a part. By a 'functional' relationship, we mean that the self model is not merely a

picture of the system, but that it is something like a small copy of the system which preserves a close resemblance with the functional relationships between various components of the entire system.[34] Most importantly, the functional relationships between components in the model may be activated with all, some, or no activation at all of the represented parts of the system. In other words, the self model needn't always mirror the exact current state of the system; the self model may become temporarily "disengaged" from the reality of the actual system of which it is a part. It is this property which allows the self model to simulate various actions of the system without requiring the whole system to participate in the exercise. Under the direction of the self model, an action may be "tried on for size" without actually engaging in it, and the information stored in the relationships of the self model may provide a rough indication of the effect on the system itself of actually performing the action. We will examine these two properties of the self model—functional representation of the system and differential activation of represented componentry of the system—each in turn.

The representation of the system itself can be understood through an analogy with cartography. One type of flat map of the surface of the Earth, called a Mercator projection, is made by a process something like placing a light source at the centre of a transparent globe on which the Earth's surface is painted. If we place a cylinder of paper over the globe, tangent to it at the equator (or, a related scheme, two cones of paper, tangent at 45° north latitude and 45° south latitude), we may trace on the paper an outline of the map painted on the globe. By increasing the diameter of the cylinder of paper, we can get a larger and larger scale representation of the map on the globe. In the case of the self model, we are doing something like the reverse process: we are shining a "light" (through what might be a very large cylinder in the map example) from the outside, through the various sensory modalities of the system, on to a representational data structure of the entire system contained inside. Significantly, the self model represents additional features which don't exist in the reverse Mercator projection example: it also represents the activity of *internal* structures. There are of course no functional

---

[34] Much later we will see how the notion of representation given here may be misleading, but for now it will do no harm to use the term, as long as we don't try to read implications into the discussion apart from what we explicitly address. (See also Chapter 11.)

relationships between curves painted on a globe (or between them and anything internal to the globe), so the analogy breaks down again there; in the self model, the "projected" ("injected"?) representation preserves the relationships which exist between components—both internal and sensory—of the system being represented. (Later we will discuss the extent to which the model may also *add* to the relationships being represented.) The self model may represent sensori-motor relationships as well as associations between sensory input or motor output and the states of other internal components.

The second property of the self model representation of the entire system of which it is itself a part concerns the activation relationship between components of the representation and components of the system represented. The key idea is that the functional relationships which exist between components of the entire system also exist between components of the representation. Moreover, activations at the self model level may remain correlated with activations at the level of the system, even if activation at the self model level has *originated* at that representational level rather than coming from system input. Thus, ignoring the system itself for a moment, we may get an indication of the response of the entire system to a given set of conditions by activating components of the self model in a way similar to that in which they *would be* activated if the system itself were actually in that set of conditions. With respect to the two-way relationship between system and self model—for the self model is again no mere epiphenomenal image—the response of the self model may even induce a response at the system level. This capacity for differential two-way activation of model components by system components and *vice versa*, allows for a complex class of overall states displaying mixtures of environmental and internal input/output activity throughout the system.

To illustrate with an implausibly simplified human example, my visual field might actually be filled (for our purposes, at the system level) with the hill and flowering trees out my window, yet I could be visually imagining (for our purposes, at the self model level) a vegetable stir-fry which causes my mouth to water. My stomach might even begin to grumble in anticipation of food at the same time that I shift in my seat because of an association with sore legs the last time I walked up the hill I see in my visual field. In this case, perhaps, the "attention" of the self

model is split between two visual images, and two different reactions are arising in the system, one as a result of direct visual input and the other as a result of imagined visual input. Here *both* the direct system level stimulation of the visual field with the hill and flowering trees and the stimulation at the self model level with the imagined visual input of a vegetable stir fry have an effect at both the self model level and the system level. We will have much more to say on this topic later; for now, however, the important point is just that correlated functional relationships may exist at both the self model level and the system level as well as between the two levels.

6.2 Compression—Fitting it All In

One feature of the self model's system representation which is present also in the reverse Mercator projection analogy is *compression* of information about the system. That is, relative to what is being represented by the self model, the actual representation is denser. Unlike the reverse Mercator projection, however, we are here dealing with discretely instantiated projections, and the compressions are what information specialists term 'lossy'—i.e., there is a loss of information as we move from the original to the compressed representation. In other words, given only the compressed representation, we cannot reconstruct precisely the source object of the representation. This is not to say that the original cannot be reconstructed at all, but only that some of the finer details may be lost. (Indeed, we shall see later that real self model representations are rarely, if ever, actually "reconstructed" at all.) What exactly we mean by this in terms of neural organisation will become clearer in our later discussions of actual architectures for implementing self models, but for now we can say that the activity of highly correlated areas of a *system* may be represented by a single smaller area of the *representation*.

This is roughly akin to a way of representing a crowd of football fans which we might use if we were only interested in how noisy they were. A single measure of increasing volume in terms of decibels or even a simpler scale of 'raucous', 'very raucous', 'obnoxiously raucous', and 'English' might serve as a lossy representation of the whole lot. From the decibel measure or from the raucousness measure, we could not

reconstruct detailed information about what each individual football fan was doing or how they sounded, but we would have an indication of their overall level of noisiness.

A key property of the self model is the capacity selectively to "decompress" this kind of representation of a given area of the system. This is not true decompression in the sense of recovering something like an original image from a smaller representation of it; instead, this decompression amounts to the self model's "looking" through the representation more closely at the actual system level activity being represented. This is best understood with a preview of the architectural implementation of the idea: when the attention of the self model is directed towards the activity of a particular part of the system available in lossy form in a representation, the output channels of the actual components of the system are disinhibited so that *they*, rather than the lossy self model representation, provide afferent signals to the rest of the relevant parts of the whole system. Thus, no representation is decoded into any "display space": the functional arrangement of the system is simply altered slightly so a different set of neurons (with the "original", more complete, information) is providing data. This is roughly akin to, say, being aware of an apple sitting on the table directly in front of you but not being able to report very much about it until you deliberately pay particular attention to extracting the details of its shape and colouring.

Of course, there is a limit to the extent of the area over which this pseudo-decompression can be carried out at a given time; the finite number of components implementing the self model places an upper bound on the number of afferent signals which can be meaningfully (in terms of functional effect on the self model data structure) processed at a given time. Such a bound on the size of the "virtual window", or "abstraction window", through which the system may be seen, however, does not by itself preclude the selective decompression of any arbitrary sequence of areas of the system in turn. This is akin to scanning an entire horizon through a pair of field glasses: although we can extract detailed information about any area of the horizon, we cannot see it all at once in such detail because we don't have enough "space" in our visual field. (The self model is lacking in "space" in the sense that it has limited componentry.) But by sweeping the field glasses across the horizon, however, we can eventually see all of it in great detail.

What may place a second limit on the self model's capacity for such selective decompression in the human, however, is the actual physical bound on dendritic arborisation, which is closely related to the number of relevant afferent signals a single neuron is capable of receiving. Although it is possible in principle that a diagram of all the connections in a brain might form what mathematicians call a connected graph[35]—that is, a diagram of nodes and connections in which it is possible to travel along some series of connections between any two arbitrary nodes—neurophysiological considerations make it highly unlikely that this is the case. Thus, there may well be a limit to the system area decompressible (even indirectly, with the intervention of a number of intermediate neurons) by any finite set of neurons implementing an area of a self model. In practice, this limit may be more theoretical than practical, since the sparse activity distributions of real networks of neurons indicates that the sheer number of neurons involved in implementing an area of a self model may generally be well in excess of the number required for reasonable decompression of any of the system areas to which it is functionally related. (We will revisit something like this kind of redundancy later when we discuss neuronal degeneracy.)

In any case, the main point for the moment is that components of the self model are lossy representations of the system, but some amount of the detail of the activity at levels lower than the representations can be extracted by looking through a "virtual window", or "abstraction window", and selectively disinhibiting the outputs of the actual components of these areas. This loss of information as we move from the system level to the self model with only selective and limited recovery of the original information is counterbalanced, however, by the emergence of higher order information within the self model, and it is to this idea that we now turn our attention.

6.3 Tricks with Information

As we noted above, through higher level representation in the self model, there is a loss of immediate information about the actual activity

---

[35] In fact, the graph would need to be *directed* and *weighted* as well, since neural signals travel in only one direction and since edges, or connections—which here correspond to synapses—have differential capacities for carrying signals (i.e., different efficacies).

of individual lower level components. This is not unlike the loss of information which goes on unavoidably in an ordinary digital computer.[36] A flip flop (occasionally called a bi-stable multivibrator or, more often, simply a bi-stable) is a lossy representation of the states of the individual transistors which make it up, just as the individual transistor states of the flip flop are lossy representations of the states of the electrons flowing through them. Subsequent logic operations lose information about the states of the individual flip flops which preceded them, and outputs are many to one logical operations on sets of these input states, so the output here is again lossy. This loss of information proceeds up through various software levels as well.

However this loss of information at each level in an actual system with a self model—as each node or neuron loses information about the individual activity patterns behind its afferent signals and outputs only its own particular spiking behaviour derived from them—is coupled with the emergence of new information at higher levels. It is at these higher levels where information-carrying correlations may be set up between nodes which are abstracting the information fed to them. In other words, the self model may lose information about the lowest level neural happenings while extracting higher level information about correlations between different parts of the system and relationships between different more detailed feature maps.

By way of analogy, we might consider the kinds of comparisons which could be drawn between trees. With very detailed information about the molecular structure of leaves, for instance, we could identify the presence of common organic molecules. By moving our observations up a level, we might notice what leaves look like, and we might be able to see the similarities between different kinds of leaves from different kinds of trees. Notice that these things don't appear if we're looking only at information about molecular structure. Moving up to a still less detailed

---

36 I am aware that it was established more than two decades ago (Bennett 1973) that computation does not *require* this kind of information loss and that a substantial literature has developed around the idea of reversible computing. This literature includes computing designs of both a classical (Fredkin and Toffoli 1982) and quantum nature (Benioff 1980, 1982; later Deutsch 1989 and others) and Bennett's (1982) own Turing machines, not to mention comments on the implications of reversible computation for the second law of thermodynamics. (See Leff and Rex 1980 for Maxwell's Daemon; Zurek 1989 for one thought on the daemon and reversible computation.) However, we are concerned with finite living biological organisms for whom the annihilation of some information—the energy costly and irreversible part of computation—is a precondition of adaptation and selection.

level, we might lose all idea about molecular structure, but we could observe that the gross shapes of many different trees are very similar to one another, and moving up still another level, we could even begin to make observations about forests. If we had some huge data collecting apparatus which could take in information about an entire stand of trees right down to the molecular level, all of these observations could be made starting with the same data set, but to avoid missing the forest for the trees, so to speak, we would need to deal with abstractions extracted from the data. That is, we would have to throw out some of the detail in the original information, representing some groups of data all together; only then could we see how the *groups* of data relate to each other.

It is just this same kind of emergent information which occurs in the digital computer—where coherent data structures can be formed out of a high level representation of a bunch of whizzing electrons whose behaviour would actually look stochastic and bizarre if we peered in on them closely enough—and which, I suggest, occurs in a system possessed of a self model. It is only at the level of representation—which for our purposes presently we can read as "level of abstraction"—available to the self model where functionally relevant information[37] can appear about relationships between various parts of the system and between the system and its environment. Indeed, in keeping with the notion that all information is and can only be physical (Landauer 1991), it is almost self evident that these relationships between various parts of a system can only emerge in a functionally relevant rôle when there is this kind of hierarchical abstraction and the loss of information which accompanies it. Next we shall explore the meaning and significance of anatomical features of reentrance and heterarchical organisation as they relate to the formation of polymodal associations from this emergent higher level information.

6.4 Associations upon Associations

The kinds of information emergence which we discussed in the previous section are of course not limited to cases where a single body of

---

[37] This qualifier is added only in view of the fact that although correlations already exist between components at the lowest levels of the system—just as the forest is already there in the mountains of molecular structure data—until they are extracted by some other component the correlations themselves can have, as such, no functional rôle.

data are being processed, as in observing forests and trees and the molecular structure of leaves. Rather, it is entirely likely that in biological systems, we will observe the extraction of information about correlations between groups of data available to two different modes of input. That is, we may see two parallel hierarchies of abstraction from two different data sets—perhaps gathered in two different modalities—coupled with the extraction of information about correlations between the two hierarchies. Something like this kind of multimodal association may occur, for instance, when objects recognised from visual cortex data are correlated with odours recognised from olfactory cortex data. Almost as if it were a degenerate example of classical conditioning, we may then observe the learned association of the visual appearance of a food with its smell.

Looking ahead again to the more detailed architectural descriptions to come, we can see that such an association might develop easily when adaptive mechanisms ensure the effective mutual exchange of excitatory neurotransmitters between two groups of neurons which are each primarily responsive to the activation of particular representations from different modalities. Where activity in the groups being represented is correlated, so, too, may the activity of the representing groups become correlated and mutually reinforcing. Even with activity inhibited from the array to which one of the groups normally responds, the excitatory input from the other group responding to its own preferred inputs may be sufficient to drive the activity of the first to significantly above the base level.

The anatomical features which enable the development of these kinds of associations are bound up with three characteristics known as *reentrance*, *recursivity*, and *heterarchy*. The first is a broad term referring to axonal arborisation from, say, a feature map, into the input areas of another feature map (or even back into the dendritic areas of the feature map from which the reentering neuron itself took its own input).[38] In other words, it is a way for two usually separate groups of neurons to share their outputs. The majority of central nervous systems in the vertebrate world are structurally reentrant, either locally or across wider areas. (Brodal 1981) Recursivity is the term used in artificial neural network design to denote the powerful technique of using feedback from the

---

38 Alternatively, the term may refer to dendritic arborisation into the appropriate efferent areas.

output layer of what would otherwise be a feedforward network back to the input layer, and it can be understood in our context as the special case of reentry noted parenthetically above. The last term, heterarchical organisation, refers simply to an hierarchical structure with any kind of interlevel feedback. Later we shall explore in more detail the rôle of these kinds of organisation—particularly of reentry in general—in enabling the development of polymodal associations in the self model. For the moment, however, there is more to say about why these associations are important at the level of the self model.

These cross-modal associations are important in the self model not only for the establishment of such basic abilities as visuo-motor coordination, but they also may well be an enabling factor in the development of language abilities. Contra Lieberman (1984, 1985, 1989), I agree with Wilkins and Wakefield (forthcoming) that the development of an "association cortex" in the so-called POT (parietal-occipital-temporal) and Broca's areas of the human brain was the anatomically necessary evolutionary precondition (which first appeared in *Homo habilis* some 2.5 to 2 million years ago) of language use, rather than the emergence of the modern vocal tract, which apparently (Arensburg, et al 1990) came much later. Wilkins and Wakefield, however, suggest that "by the time sensory information is transferred to the POT, it has already undergone higher order processing and relinquished its uni-modal character; POT representation of sensory input is therefore entirely modality-non-specific, or modality-free". (See Pandya and Yeterian 1985 for more specific cytoarchitectonic features of the POT which bear on its integration of data from the three primary neocortical sensory association areas.) The conclusion of this sentence appears to be a *non sequitur*, and the conclusion in general is, I believe, deeply misguided. The cytoarchitectonic properties of the POT to which Wilkins and Wakefield point do suggest the emergence of *polymodal* associations, but getting from there to amodality is another matter.

Unfortunately, I do not have the opportunity for an extended engagement on this issue, but the basic point can be made very easily: Wilkins and Wakefield give too little consideration to the fact that neurons always get their inputs from *somewhere*. Their theory of association in the POT and Broca's areas is meant to be compatible with Jackendoff's theory of conceptual structure (see, for instance, Jackendoff

1983, 1987, 1990), but I believe the anatomical price is too high for the kind of rendition Wilkins and Wakefield are seeking. Specifically, we should not give up the anatomical *grounding* (for entry into the growing literature on the symbol grounding problem, see Harnad 1987, 1990, 1992, 1993; Harnad, et al 1991; also Ford and Hayes 1991) of association symbols in *polymodal* data for the sake of achieving abstraction.

That is, I believe it is a grave error to suggest that representations in a real biological organism (as opposed to, say, representations in a book) could be abstracted completely away from *some* kind of modal input. This is to say that any abstract representation must be applicable to some kind of (perhaps polymodal) input and, moreover, that it must similarly have *arisen* from such input. To put it still another way, a representation may get farther and farther away from the sensory inputs, but it cannot exist in a functionally relevant rôle in complete isolation from the world.[39]

We will examine later the possible rôle of representations of the self model itself, a purely internal set of structures, yet even in this case—insofar as the self model itself is a model of a system and its sensory windows on the environment—any purported complete divorce from polymodal input remains an illusion. Indeed, it may be the very efficacy of polymodal representations in language itself and the relative strength of associations *between* representations partly abstracted from polymodal data as compared to the strength of the associations between the representations and the polymodal data themselves which is responsible for this illusion of amodality. For the moment, I leave the idea as speculation, but there may be considerable mileage in the notion that this illusion of amodality is even at work in persuading us of the *prima facie* plausibility of the general existence of non-relational qualities of experience, or qualia. The same kind of illusion might occur with respect to abstractions arising from a *single* modality; thus our capacity to *name* some abstraction common to all varieties of a particular kind of modal experience but allegedly independent of the particular relational qualities of any of them (the "redness" of the red, for instance) may lead us to

---

[39] Lest this be confused with the incorrect assertion that signals from a representation might somehow flow "backwards" to those groups which activated it, notice that we are saying only that representation is *responsive to* polymodal activations; this doesn't imply that activation of a polymodal representation need directly cause the activation of anything in neural assemblies dedicated to those modalities which activate *it*.

believe that such associative abstractions accurately reflect real properties of our experience.

In the present context, however, we must leave such interesting questions to the side and continue with our tour of basic properties of self model data structures. Before turning to features of self models related more to their interactions with the environment, we next consider another property closely related to the emergence of the kinds of higher level information and polymodal associations which we have just been exploring.

6.5 Building on Associations upon Associations

This property derives again from the emergence of higher level information and the extraction of polymodal associations by the self model. Specifically, we are concerned with the capacity for applying new (i.e., not previously experienced) combinations of associations which this enables. That is, we should expect in the self model a capacity for "connecting" previously independent areas of activity and for building new correlations where before there may have been none. This robust combinatorial capacity is of course significant for a variety of endeavours, but arguably one of the most important and sophisticated of them is in the kind of linguistic application of polymodal associations which we discussed above.

In the first instance, this capacity will be limited in the human case by the dendritic arborisation of candidate neuronal groups (or sets of groups) which may be available to represent a particular combination of other groups. That is, in the first instance, we should expect there to be no immediate combinatorial capacity where we find an empty set as the intersection of the sets of neurons taking afferent signals from those neuronal groups active in the representations we might want to combine. However, this limit is mitigated both by the capacity of biological networks to grow *new* connections and—recalling the previous note on heterarchical organisation—by the fact that a given neuronal group may be active in two or more distinct representations occurring at two different levels of abstraction. The rôle of the first capacity is obvious; the significance of the second is that even if there are no *direct* connections between sets of neurons taking afferent signals from different groups,

there may be interlevel connections which could ultimately be strengthened enough that the functional outcome would be similar to or identical to the straightforward availability of the kind of "linking" group indicated as a non-empty intersection of relevantly input-connected groups.

Likewise, new correlations might be extracted when two neuronal groups whose afferent signals come primarily from widely separated areas are themselves physically near enough to form reentrant connections. Without the hierarchical organisation which provides these two proximal groups in the first lace, of course, association between physically separated processing groups might never occur. To simplify greatly, the deeper the hierarchies in place for any two areas of processing, the greater the potential for forming complex associations between the two.

We will return to this last idea when we discuss selective advantages of self models, but for now with these limits and mitigating factors in mind, we can see that whatever area is responsible for the emergence of polymodal associations active in language use must be highly structurally reentrant, and the dendritic arborisations of distinct neuronal groups must often overlap a great deal. Both the capacities we have described for the linguistic application of polymodal associations and the limited conclusions about the cytoarchitectonic features of an area capable of this kind of associative integration are consistent with the neuroscientific data. (Geschwind 1964, 1965; Pandya and Yeterian 1985; Greenfield 1992) This is especially true with respect to the input to the parietal-occipital-temporal area (and, obviously, its internal connectedness), which integrates information from the three neocortical association areas dedicated to audio-visual convergence, audio-somatic convergence, and visual-somatic convergence. (Pandya and Yeterian 1985; see also Ingvar 1985 for comments on integration and temporal processing in the prefrontal cortex.) These observations are also consistent with data (Humphrey, et al 1979; Wagner, et al 1981) concerning the "metaphorical mapping" capacity of human infants to form concepts more abstract than the data on which they are based. Indeed, given this kind of capacity, Wilkins and Wakefield suggest that it is specifically the POT and Broca's areas which are not only a necessary precondition of language acquisition but which are uniquely responsible for the general ability to abstract properties. It would be premature to jump in with Wilkins and

Wakefield and conclude, essentially, that the association abilities of the POT and Broca's areas, coupled with certain evolutionary changes in the premotor cortex, are both necessary *and* apparently sufficient for language acquisition, but this kind of position at least serves to illustrate the importance for linguistic abilities of the kind of combinatorial capacity we are addressing.

Yet again, however, we must leave further exploration of this area for another time and turn to what is perhaps the less exciting feature foreshadowed previously by the comment on symbol grounding: the rôle of environmental feedback loops for self models in real biological systems forced to compete for survival in the world.

6.6 Feedback

For the context of our present explorations, we take it as given that the original evolutionary purpose of a mind—and, thus, the purpose of a self model—is to help an organism get on in the world. Thus, whatever areas of the brain or other self model substrate may be dedicated to higher order processing of polymodal associations, such processing demands first the direction of enormous resources towards basic mapping of the environment and the relationship between the environment and the organism as they both change through time. Apart from the sets of sensory receptor sheets serving each modality which act as the first stage for feature extraction, this requires in the first instance tuning of the sensory systems themselves. This depends upon close communication between the systems which control the physical states of the organism's sense organs—or at least those which control the states which make a systematic difference to the efferent signals from the relevant sensory systems—and those systems which take their afferent signals from them. (In the human, these sensory systems are primarily but not exclusively controlled by the motor cortex.) That is, the system must use environmental feedback to develop *coordination* between sensory and motor systems.[40]

---

[40] Here and elsewhere, it is important to bear in mind that feedback may be either positive or negative. In the first case, feedback helps to select by reinforcement those neural groups which contribute to desirable or efficacious actions, increasing the chance of those groups firing on the next relevant occasion. (At a higher level of organisation, this is something like learning that eating chocolate is pleasurable.) In the second case, feedback

In the human, for instance, we require close connectivity between the occulomotor system and at least the primary visual cortex, if not most of the rest of the visual processing system as well. The reentrance between the two systems allows the two to work together in a coordinated fashion for the purpose of providing data to the rest of the organism. (Of course, the reentrance needn't be direct; such coordination may also be subserved by feedback mediated by other processing arrays, as in the second example network of Chapter 10.) This general observation is consistent with anatomical data, which indicate the relevant architectural features in callosal connections, thalamocortical and corticothalamic radiations, and various other links between sensory and motor areas. (Zeki 1975, 1978; van Essen 1985) Contra Edelman, however (see Chapter 9), who makes much of the requirement that reentrant signalling of this type be *phasic*, I suggest it is not necessary that the activity of the connected maps be temporally linked in a linear fashion. Instead, all that should be necessary is that there be a continuous and consistent functional relationship (perhaps even a nonlinear one, nonphasic but also not chaotic) between the activity of linked areas.[41]

Such basic bimodal coordination, however, does not by any means exhaust the rôle of environmental feedback in the self model. There are two primary reasons why this is only the beginning. First, it is clear that the organism requires much more complex global mapping between two *or more* modalities to perform even the most basic tasks. For instance, reaching out to grasp an object requires, among other things, complex mappings of relationships between motor output and visual input (something like hand-eye coordination, necessary so the object may be brought within the organism's grasp), between motor output and other somatosensory input (something like "knowing one's strength", necessary so the object may be grasped but neither crushed nor dropped), and even

---

helps to damp out the activity of neuronal groups which contribute to an undesirable or ineffective action, making those groups less likely to fire on the next similar occasion. (At a higher level of organisation, this is something like learning that touching a hot iron is liable to make a painful burn on one's hand.)

[41] This is because there is not, at least *prima facie*, any reason to suppose that the system could not still function even if the relationship between any two arbitrary systems requiring coordination was altered according to almost any continuous function. (By this we may mean not only that the relationship is phase shifted, but that it might even be phase shifted by amounts varying according to the underlying activity.) As long as the function may be learned by a neural network, there is every reason to think that a human system, for instance, ought to adapt to such an alteration in somatic time.

between visual and somatosensory input (which enables the system to "know" that it is indeed its own hand, for instance, which is grasping an appropriately shaped object which it sees in its visual field). We may expect that at the level of the *self model*, these mappings must typically incorporate all sensory modalities plus the motor systems so that the model of the organism is a completely polymodal one with continuous relationships between a continuous range of possible combinations of sensory input and motor output. That is, the model of the self is not a model of what it looks like or how it may move; it is a complete model of (among other things!) how its movements and sensory modalities are all related.

The second reason we've really only scratched the surface of environmental feedback in our discussion so far is closely related to this last point. Specifically, the basic low level bi- or trimodal coordinations we've posited are unlikely to feature in the higher level self model. That is, the self model itself is unlikely to have "direct access" to the information establishing this coordination and is likely instead simply to call on it through something akin to the "abstraction windows", or "virtual windows" we discussed earlier. In other words, actions might be initiated at the level of the self model which require this kind of low level coordination to be in place, but for the most part we should expect this coordination to be *assumed* by the self model (once it has emerged developmentally and epigenetically, of course). Normal adults, for instance, are capable of reaching out to grasp an object, but it seems likely they could not immediately provide any information at all to describe the specifics of how their various visual and somatosensory feedback was related to their motor actions. They might be able to *construct* such descriptions through conscious effort—something like looking through the abstraction windows in a systematic way to learn at a higher level something about the mappings already in place at a lower level—but we should expect the actual connections between the relevant sensory receptor sheets and so forth to be at a much lower level than that of the self model.

With this in mind, then, we can see that the primary importance of environmental feedback for the self model is in the availability of systems to *subserve* the self model's need to model the interaction of the system with its environment. That the self model even requires such systems is

actually a substantive statement about the minimum requirements for being a *self*. Expanding on the notion with which we opened the discussion of environmental feedback, we here take the position that a self is a *situated* self whose activities, pace Harnad, are *grounded* in a real world. Motivation for such a view comes from common sense about our own conscious existence as well as from both evolutionary imperatives—selves developed under pressure to survive in an environment—and general considerations in the philosophy of mind which have sought to replace naïve mechanistic cognitive scientific accounts with more advanced teleofunctional frameworks. Here is unfortunately not the place for a discussion of all the issues surrounding the context in which a self may be situated, but it is worth noting that the view we have touched upon here is particularly complementary with the position taken on meaning by the likes of Putnam (1975, 1988) and Burge (1979, 1982, 1986)[42]. (Also see Stalnaker 1993 especially.) We will return to this matter briefly when we turn to consider more carefully what we mean by *change* in a self model.

Having explored the rôle of environmental feedback, the last of the basic properties of self models we will examine for the moment, we now move on to some brief comments on the relationship of the self model to functionalism. This is followed by a quick rundown on the selective advantages of organisms equipped with self models and the evolutionary emergence of consciousness. Finally, we move on to somewhat more technical discussions of embedding self model data structures in real neural networks.

---

[42] I would certainly want to distance myself from some of the stronger conclusions drawn by these authors and their commentators, yet it is interesting to notice the complementarity at this early stage, before other agendas drive the overall views apart.

7

Functional Selves

This view we have been outlining, which relates sensation to changes in an instantiated *data structure* called a self model, is complementary to a functionalist approach to cognition but is by no means identical to it. Above we have occasionally referred to functional relevance *relative to* the self model, but this is different to functional relevance in the system itself; it is now time to clarify what we mean by this and to pin down more carefully how self models are related to functionalism.

7.1 Hunting Self Models

First, it is important to be clear on how we discern the self model data structure and distinguish it from other data structures which might be implemented by the same material structure. Notice that the same material structure may be interpreted from the outside—that is, from the third person point of view—as instantiating any number of different functional relationships and data structures. This is trivial: for any functional description of either the relationship between different material components of a system or the relationship between different bodies of data in a data structure, we can conjoin to this description an infinite class of additional functional descriptions of counterfactual conditions which in practice fall outside the domain of the system in question. This is analogous to the observation that there is an infinite class of functions which pass through any finite set of points.

Someone might object that we could pare down the class of functional descriptions of a material system or of the data structures it might be instantiating by providing our own class of counterfactuals covering areas outside the domain of the system in question. That is, we might say that given some state the system in which the system will never

be, *if* the system *were* in that state, it *would* evolve exactly thus and so. But this is unsatisfactory for a number of reasons, not the least of which being that this absurdly makes understanding the experience of a self model operating over a given domain contingent on the provision of descriptions of behaviour otherwise utterly irrelevant to the self model. This is akin to providing many other points to restrict the class of possible functions describing a curve through some small set of points in which we are actually interested. Perhaps more importantly, such logical jumping through hoops is simply unnecessary. The relevant class of data structures (or functional descriptions) can be extracted from the infinite class by appeal to the system itself.

That is, we are concerned with the data structure implemented with respect to the structure *itself*; we are interested in the data structure incorporating a coherent high level model of the system itself and which displays the kinds of properties we have previously been outlining. That there may be more than one such data structure which could be attributed to a given material arrangement is, for our purposes, irrelevant. This is simply a fact of life in trying to describe all interactions of material things: even the laws of quantum mechanics are not the sole functional descriptions of all hitherto observed phenomena! We merely choose the most parsimonious descriptions which cover the relevant domain and leave it at that. It is no different for the self model: we simply choose a functional description with respect to the system itself and which fits the domain in which the system may find itself and stop there. As long as we understand how the model changes in that domain, we have satisfied our needs.

We might notice, incidentally, that none of these observations are in any way incompatible with an entirely deterministic material implementation of the self model. The self model is an abstract data structure, and—unlike the simple stack example, in which the data in the stack can be read off from the states of microcircuits—it may include in itself functional descriptions of how parts of itself relate to each other or of how parts of the system relate to each other (in terms of how the organism uses it, this is indeed largely the point of the self model!), and these functional descriptions can be just as plural in broader (and irrelevant) domains as any scientific law might be.

7.2 Hunting Functional Relevance

Having noted the subtlety of relating data structures to the functional arrangements instantiating them, we can now say something about functional relevance relative to the self model. This idea becomes clear as we pin down the relationship between mental experience as viewed as changes to a data structure and mental experience as viewed as the existence of a material structure in a particular functional state. Although there are many things to be said on this topic, perhaps the most significant is that we really do mean to link mental experience to *change* in an instantiated data structure and not just to causally efficacious states in some functional assembly. That is, sensation is a *process* of change in a data structure; it is not merely the *existence* of a particular data structure. Where there isn't this change, there isn't experience: thus if we were to "freeze" a self model in time, it would cease to be a subject of experience.

On this view, talk of a "mental state" makes sense only in the context of a temporally evolving self model data structure. A "mental state" in this context is something like the derivative of a function describing a curve at a particular point: there is an instantaneous rate of change for a function only because there is some quantity undergoing change. A point on a curve has a slope only because it is a part of a curve. Likewise, there is an instantaneous "mental state" only because there is a changing data structure. This is in contrast to the functional view of a mental state, which corresponds to a functional state. The difference is that for the functionalist a static functional state still apparently qualifies as a mental state, whereas on the self model view there is no such thing as a static state with qualities of experience.

Another consequence of the self model view is that if a self model is being instantiated by a particular functional substrate which is in fact *changing*, perhaps in order to, for instance, maintain synchrony with a changing environment, but if the data structure is rendered static by this *functional* change, there is no experience. This relationship is simple to quantify: the functional system may change without changing the self model data structure and thus without giving rise to any conscious experience, but the data structure may not change—and thus conscious experience cannot occur—without change in the functional system. Thus our use above of the notion of changes which were functionally relevant

to the self model was meant to pick out those changes both functionally relevant to the system and capable of bringing about a change in the self model data structure.

7.3 Self-Centred Change

There are two more points to explore with respect to change in the self model before we move on to evolutionary considerations. One is rather simple but carries significant implications, while the second is more complex but is perhaps not as significant. The first concerns what we mean by change in the data structure. It is simply this: by *change*, we mean change relative to itself. While it is a basic point, this fact actually has powerful consequences. Recall from the discussion of environmental feedback that the view on offer considers selves in an environmental context in which their actions and their internal representations are grounded. We noted that this is nicely complementary to points made by Putnam, Burge, and others. However, it is here that we must be very careful about how much we read into this complementarity. Because we are linking sensation to change in a self model relative to itself, we are divorcing experience from changes in the environment which don't bring about such change in the self model relative to itself. More to the point, changes in the environment which bring about a change in the relationship of the self model to that environment but which do not effect a subsequent change in the relationship of the self model to itself are irrelevant to the experience of that self model. We can see this more clearly through another way of looking at Putnam's own Twin Earth example.

The point of the original example was that two otherwise identical people, both thinking about "water", could each be referring to different things, on account of the fact that one of them lives on Earth, where water is H_2O, and another of them lives on Twin Earth, where water ("twin water") is actually something else. Therefore, reference—and meaning—can't be just "in the head". Now on one way of viewing the twin people as compared to the people, a way presumably more friendly to Putnam's preferences, their self model data structures must actually be different—since the latter contains data about water and the former data about twin water—even if their *physical* structures are very similar or even identical.

When we consider any single self model on this view, we can see that any change in the world which "shifts" the reference of some element of data in the self model actually brings about a change in the data structure. (This would be akin to, for example, the world changing in such a way that all water became twin water, thereby changing all people into twin people.[43])

But on the present view, the self model is a data structure defined with respect to the *system*, and change in the self model refers to change with respect to *itself*. In information theoretic terms, the self model data exists in correlations between states of whatever is instantiating it and states of the system as a whole (or, more precisely, correlations between functional relationships in whatever is instantiating it and functional relationships in the system being modelled). While large parts of what is being modelled in the self model concern the relationship of the system to its environment, that relationship is always understood through the sensory inputs to the system. That is, the data structure models the environmental interactions of the system by modelling the interactions of the system and the system's sensory inputs. Thus, a change in the environment which does not bring about a change in the system's sensory inputs (or a change in how those inputs relate to the activity of the system) does not bring about any change in the self model *relative to itself*. It has no effect on sensation.

Thus, the self solipsistically survives without change any changes in the environment which have no functionally relevant effect on it. The *meaning* of representations in the self model, as understood from the outside, the objective point of view, may be changed without there being any change whatsoever to the rôle of those representations "on the inside", from the subjective point of view. Meaning, then, may be different between the objective and subjective perspectives. Of course, none of this is to say that there mightn't be some great change in the self model when it learns that there has been a shift in the world which has

---

[43] This is a debatable point, whether such a change from water to twin water would in fact change all people into twin people, since arguments might be mustered to the effect that if all the (twin?) people could somehow be informed that the reference of their word 'water' had actually changed to some other substance than they thought, then they could object that they actually meant the other kind of water. But the mechanics of this particular change do not concern us so much: we are only seeking to clarify the simpler point that there may occur changes in the world which alter the "meaning" of data in the data structure but which still do not count, on this view, as bringing about a change in the data structure.

changed the reference of some one or more of its representations, but of course this *learning* of the shift amounts to a change in the self model data structure; this lack of awareness before we have learnt about something coupled with awareness when we have learnt about it is obviously exactly what we should expect. That it requires us to deviate from what the likes of Putnam might prefer is perhaps a criticism of Putnam's overall approach, but it is no criticism of the present view, which remains entirely in line with common sense.

7.4 Good Vibrations (Oscillations)

The second of the two final points to be explored on the topic of functionalism and change in the self model concerns implications for the kind of low level dynamics in the brain (or whatever) this view requires to support continuous sensation. Recall that above we distanced ourselves from the functionalist's equation of mental state with functional state and noted instead that the idea of a mental state at a given moment makes sense only in the context of a continuously evolving self model data structure. A straightforward implication of this view is that if it is possible for a self model to exist in a state of unchanging sensation for more than an instant of time, then that state must actually correspond to some kind of continuous *change* in the self model. In the apparent absence of convincing reasons why it should *not* be possible to experience such enduring states of sensation[44], it is worthwhile considering the relationship between a continuously changing data structure and unchanging sensation.

We begin by noting that rich neurophysiological evidence indicates oscillatory spike frequency activity in cortical areas associated with the recognition of particular stimuli presented to an organism. Decades of extensive research by Freeman and his colleagues (Freeman 1964, 1972, 1975, 1979, 1987a, 1987b, 1988, 1989, 1991a, 1991b; Freeman and Skarda 1985; Skarda and Freeman 1987; Yao and Freeman 1990; Eeckman and Freeman 1991) into the olfactory system, for instance, indicate not only the

---

[44] Notice, however, that the reasons why we *should* be able to experience enduring sensation are far from obvious: after all, most of the time we are experiencing some continuous sensation in one or more modalities, our mental state is still apparently changing, as we silently talk to ourselves or subtly shift the sensory receptors giving us the sensation or even as we become aware of the passage of time.

importance of the overall dynamical character of large collections of neurons in recognising odours (as distinct from the particular behaviour of individual neurons or of smaller groups of them) but also the preponderance of theta (roughly 5 Hz) and gamma (roughly 40-70 Hz) local field potential oscillations. Many suggestions have been offered as to the functional rôle of such oscillations, including improving the exchange and recall of information between the primary olfactory bulb and the olfactory cortex. (Ambros-Ingerson, et al 1990) Others (Li and Hopfield 1989) have proposed that oscillations may help amplify weak signals and enhance response time, while Wilson and Bower (1989) believe oscillations in feedback loops could help a system compare afferent signal patterns with stored patterns. Gamma rhythms are proving an increasingly popular object of study, especially in experiments akin to those of Gray, et al (1989) which indicate the apparently stimulus induced appearance of such oscillations in the visual cortex of the cat. They have fuelled the development of accounts by Nobel laureate Francis Crick and others which apply the oscillations to the so called "binding problem" (related to Harnad's symbol grounding problem) or even to the emergence of consciousness itself. (Crick 1984, 1994; Crick and Koch 1990, 1992) Data also indicate that abnormally low availability of neuromodulators such as acetylcholine (and/or an overabundance of acetylcholinesterase?) is correlated with memory impairment in Alzheimer's disease. (See Liljenström and Hasselmo 1992 for one entry into the relevant literature.)

The robust evidence concerning the presence of such cortical oscillations suggests they are not merely epiphenomenal artefacts of neural dynamics but are instead a functionally relevant emergent feature of the collective activity of large groups of dynamically correlated neurons. (See Wright, et al 1993, for instance.) Their rôle in the present context, I suggest—admittedly, rather speculatively—may be in effecting the kind of change in the self model data structure which is necessary for continuous sensation. That is, such oscillatory activity may bring about a cyclic change in the functional relationship between components of the data structure such that continuous sensation may emerge. This suggests, again along the lines of our original observations about functionalist renditions of mental state, that a given quality of continuous sensation does not correspond to the existence of the data structure in a particular state but that it corresponds instead to cyclic activity through a continuous set of

closely related states. (On this view, it is useful to note, the sensation of recognising a particular pattern of stimuli changing in relation to a background pattern corresponds to a cyclic change in a part of the data structure which is gradually changing with respect to other parts of the data structure; this, in turn, corresponds to quasiperiodic activity at the neural level as the location in phase space of what would otherwise be, for instance, a stable limit cycle, is gradually shifted.[45]) We can see that on this account, mental states may correspond to something like attractors in an abstract phase space describing the time evolution of the self model data structure, rather than corresponding to points or neighbourhoods in such a phase space. Of course, there needn't be a straightforward relationship between dynamics at this level and dynamics at the actual neural level. (We will discuss such points in more detail in the second half of this dissertation.)

As we noted earlier, this final observation about change in the self model is somewhat more complex than the first point about the notion of change being always relative to the self model itself, but perhaps in the end it is not quite as significant. It is, after all, in line with what we might expect: the view of a continuous mental state corresponding to some sort of static functional state is somewhat naïve, and it is rather unremarkable to suggest that there is still *something* changing in a systematic way "behind the scenes" to enable continuous sensation to take place. That we have maintained compatibility with the neurophysiological data is likewise a significant but unsurprising feature of the self model view: on the naïve functionalist view, we would have either to deny the possibility of continuous sensation or to deny the functional relevance of cortical oscillations in any neural process which could give rise to sensation, thereby uncoupling the subjective character of sensation and the functional state of the instantiating structures.

With these observations to hand about the relationship between functionalism and the self model view, we may now move on to consider conscious sensation from an evolutionary perspective.

---

[45] For instance, we might observe a robust gamma oscillation whose phase is continually shifting. We might graphically depict something similar to this with a wheel skidding along a horizontal surface: the wheel might spin at a fairly constant speed, with all its different parts maintaining the same relationships to each other, while the centre of the wheel gradually shifts along horizontally. Such long term quasi-periodic oscillation could be mistaken for the aperiodicity of chaotic dynamics on a strange attractor.

8

The Evolving Self

For the present context, we shall accept without argument the evolutionary biology originally inspired by Charles Darwin (1859, 1872) as the framework into which our understanding of the development of self models and consciousness must fit. No philosophical theory of consciousness or cognition should proceed without an awareness of the biological realities in which conscious organisms live. However we understand the emergence of the self model, it must confer on the organism one or several advantages which improve its chances of producing viable offspring and maintaining or increasing the representation of its genes in future generations. If they fail this test, while self models might be an interesting exercise in theoretical cognitive science, they can hardly be a biologically plausible account of the emergence of sensation.

It is not our project to show either that the self model is the only possible architectural development which could have endowed organisms with the kinds of selective advantages we shall discuss or, alternatively, that the self model is even the most likely or most advantageous architecture which might have evolved. We shall merely explore whether the self model does indeed offer appropriate selective advantages; in the absence of convincing biological reasons why the self model architecture should *not* have developed (because the architecture is incompatible with what may be instantiated by the neural wetware or whatever other reason), we shall take this as establishing, at least *prima facie*, the requisite degree of biological plausibility for self models.

8.1 Who Needs a Self Model?

In terms of improving behavioural efficacy and survivability—and, thus, selective advantage—the self model has much to offer. We will here

concentrate on only the simplest but also the most significant of the relevant self model features. The most important feature is of course the self model's capacity to model possible repertoires of behaviour in advance of actually performing them and the kind of sophisticated deliberation over the best course of action in a real or imagined situation for which this is arguably a precondition. This capacity carries important implications for communication between organisms and the development of social living. Also significant is the improvement in response time made possible by the fact that the self model need deal primarily just with "compressed" representations of more elaborate systems, without losing the ability to call on those more elaborate systems when their full capabilities must be exploited. Closely related to this feature is the ability of the self model to take high level "short cuts" which bypass slower but more complete low level processing. Finally, since there is in real biological systems an actual physical limit to axonal and dendritic arborisation, use of a self model allows the development of more and more complex relational structures while maintaining a coherent higher level organisation capable of taking advantage of them. This also enables a kind of "top-down" learning in which skills might be learnt between compressed representations at the self model level and then "pushed down" into lower level connections. We shall discuss each of these features in turn. First, however, let's take a brief detour to note a few points we should keep in mind about applying the evolutionary framework in this context or in any other.

8.1.1 Before the Chicken and Before the Egg

Most importantly, while the information that may be preserved about successful individuals in a population is passed on through the genome, and the gene is arguably the fundamental unit of selection, we must remember that in a sense natural selection *acts* on the *phenotype* and not on the genotype. This is for the simple reason that it is phenotypes who do the reproducing in a population: organisms do not spring fully formed from a genotype, in some biological equivalent of Athena's emergence from the head of Zeus. Ontogenetic development is, then, just as much a part of the evolutionary picture as lower level phylogenetic change. Thus, understanding fully the genetic distribution in a population at a particular time requires understanding the interaction between, on the one hand, the low level mechanisms of crossover or

recombination and mutation and, on the other hand, developmental and epigenetic processes which take place in somatic time and which are determined by a *combination* of genotype preconditions and environmental interaction. In what follows, we shall leave to the side the lower level questions and devote our attention to the part of the picture concerned with ontogenetic change and the selective advantages conferred on an adult organism by the emergence of the self model. Again, this is not to say that the developmental and epigenetic processes responsible for the appearance of the self model are not in large measure *enabled* by (or even partly "hard wired" by) the genome—far from it—but, as Changeaux and colleagues point out (Changeaux and Danchin 1976; Changeaux, et al 1984), at least in the case of humans the complexity of the genome is insufficient anyway to account for the connectional complexity of the nervous system of an adult of reproducing age. With these basic observations safely stowed in the background, we continue with our discussion of the relevant advantages of self models.

8.2 Language and the Self's Self Modelling

As we noted above, the most significant capacity enabled by the self model is the modelling of possible behaviours in advance of physically engaging in them. Consider a less developed organism capable of reinforcement learning but devoid of a self model. In response to each change in the environment, the organism's behaviour is moulded primarily by genetically determined tendencies and reactions learnt from previous encounters. While there might be competition within the system between two or more different possible behaviours, each of which might have proven similarly favourable in other similar situations, in the end the organism simply responds with one of the available behaviours (or perhaps some combination of them). At first blush, anyway, there is no reason to suspect that the organism has undertaken any kind of sophisticated comparison of the behavioural repertoires apart from some comparison of what has worked for it *in the past*. Contrast this with an organism equipped with a self model, however, where we may expect not only a comparison of the past behaviours but perhaps even an actual "imaginative" simulation of possible behaviours resulting in predictions about how this or that behaviour may work *in the future* with respect to a

new and perhaps previously unexperienced situation. Of course, these kinds of descriptions needn't imply any sort of computationalism: comparisons and predictions may be implicit, and, as Edelman might be at pains to point out, they may derive from very non-computational instantiating structures.

The self model accomplishes this feat by using, among other things, its fundamental feature: a complete polymodal representation of the entire system and its relationship to the environment. By way of the polymodal associations linking motor actions of the entire system and the sensory features of the environment, the self model might, without activating actual motor units or receiving actual sensory input corresponding to a given situation, "imagine" the outcome of a particular behaviour in a given situation. While the information which enables this sophisticated modelling is based on the previous experience of the system, it in effect allows predictions to be made about combinations of sensory input and motor output not necessarily experienced previously. (This predictive capacity relies just on the continuity of the various sensori-motor relationships.) Significantly, this "open" nature of the simulation process may result in the selection of a novel behaviour with which the system does not have any actual (i.e., non-simulated) direct experience.

Given the fact that environmental situations may change enormously over the lifetime of a single individual and certainly over the course of phylogenetic development of a species, and given that information coded in the genome cannot (save accidentally) anticipate all the possible novel environments in which a phenotype may find itself, the immediate and direct survival value of this capacity can hardly be understated.[46] Any organism with the capacity to compare various possible responses to a new environmental situation must, *ceteris paribus*, have a considerable selective advantage over a similar organism lacking such a capacity. But the advantages of this capacity extend beyond the direct improvement in response to novel situations immediately at hand.

The ability to model the self in different situations without actually being in those situations may also be crucial to the development of the kind of highly flexible linguistic communication which features in the

---

[46] This capacity might also allow an organism to "practice" a particular behaviour without actually performing it, allowing mental "rehearsal" during safe times, for instance, to improve its response when next threatened.

social interactions of higher mammals such as humans.[47] I wouldn't want to suggest, as some have (Barlow 1980, Humphrey 1984), that the very *purpose* of consciousness is to enable social communication. But it is certainly biologically plausible to suggest—here following roughly the line of Wilkins and Wakefield discussed earlier, with respect to the POT area and the development of "amodal" representation coupled with premotor changes in response to selective demands on manual dexterity—that the architectural preconditions for the development of the self model's polymodal mappings of organism and environment interactions were *appropriated* for use in the kind of polymodal linguistic representations essential for such complex communication.

Moreover—and more interestingly, in our context—the modelling ability of an organism with a rich self model data structure may enable the organism to conjecture about the mental state of another organism by "imagining" itself in an environmental situation similar to that in which it observes the other organism to be. In other words, the self model equipped organism may place itself "in the other's shoes", so to speak, and so perhaps come to "empathise" with the other organism. Unless language is genetically hard wired, as it might well be in the case of social insects, for instance (see footnote preceding), this capacity to empathise would surely be of no small help in the development of symbolic language. An organism without a self model might, for instance, on repeated contact with another organism uttering some particular sound or making some particular gesture, come to associate that utterance or gesture with some kind of pending behaviour on the part of the other organism or even to correlate it with some other environmental factor. But an organism with a self model might learn to correlate the utterance or gesture not only with observed changes in behaviour or environmental features from the "outside" point of view but also with mental changes modelled from the "inside" point of view. That is, the organism could correlate the utterance or gesture with the experience *it* might be having if it were in an environmental situation similar to the one in which it observes the *other* organism.

---

47 Here I mean to distinguish human-style flexible linguistic communication with astronomical capacity for the formation of meaningful symbol strings from the rigid communication systems of, say, the social insects.

Insofar as a capacity for complex communication is an advantage to an individual, and insofar as complex communication is a precondition for the kind of social living observed in the higher mammals (and insofar as *that* is advantageous for the individuals involved), this furthers the case for the selective advantage of organisms equipped with a self model. The case is even stronger when we consider the species advantages central to population biology's phylogenetic emphasis.

Finally, we might also wonder, more speculatively, about the purely internal advantages for an organism which has developed this kind of symbolic language capacity. Given the noted relationship between the emergence of sophisticated polymodal associations and the development of symbolic language, perhaps a symbolic linguistic capacity could improve an organism's ability to "remember" low level associations at a higher level. In particular, high level linguistic representations of lower level cross modal relationships might be "rehearsed" and committed to memory more efficiently, since "learning" a relationship with a higher level lossy linguistic representation might require the modification of far fewer neural connections than learning the same at a lower level. If this is so, we might expect a linguistically competent organism to engage for a significant amount of time in constructing linguistic symbol strings purely for its own benefit as a means of raising up the contents of lower level associations into the higher self model level (where they might be analysed or just committed to higher level procedural memory): we would expect the organism to "talk to itself". That humans, anyway, *do* arguably spend a huge amount of time with this internal dialogue of course doesn't imply that this view is correct, but that it appears to be a likely consequent of the view on offer at least provides an encouraging compatibility between informed speculation and introspectively available evidence.[48]

As an aside, we should note that an organism's possession of a self model does not necessarily imply that it should develop symbolic language in an appropriate setting. Many other factors may be involved, including the complexity of the self representation itself, the availability of other "information management" faculties for the cross association of linguistic symbols each with other linguistic symbols as well as with other

---

[48] As usual, citing evidence which "tends to confirm" a theoretical account smacks of the logical fallacy of affirming the consequent!

sets of polymodal associations (which needn't necessarily be linguistic), the structure of procedural or episodic memory and its relationship to these features for cross association, the development of appropriate motor control structures for systematic auditory or manual expression, environmental resource distribution suited to sufficiently high probability of contact with other organisms, and so on. Accordingly, if possession of a self model does not imply the emergence of symbolic language, then neither does absence of symbolic language imply the absence of a self model or the absence of conscious sensation. Evolutionary pressures may continue over time to guide phylogenetic adaptation in the direction of cytoarchitectonic features conducive to the development of symbolic communication, but not all of the preconditions of such development need by any means appear simultaneously. That said, we may move on to other selective advantages not directly linked either to language or to modelling of the organism in novel situations.

8.3 Time is Survival

One of these advantages is the improvement in response time for an organism directed by a self model's "compressed" versions of its various systems. An important feature of such data structures is the "compression" of each of several interacting systems into smaller representations, as in the reverse Mercator projection example. Unlike the mapping example, however, these compressed representations maintain their functional relationships to other compressed representations. Thus, the consequences of the activity of a particular low level system on one or several other low level systems may be rapidly "calculated" in lossy fashion at the higher level of the self model before direct signals from the first low level system could have propagated to and been processed by all the others for which it provides afferent signals. This is particularly useful where, for instance, there exists a large propagation delay between the availability of output from one system and the emergence of output in response from some other system connected to it.

This is closely related to the idea of a high level architectural "short cut" which may bypass slower but more complete lower level processing. For example, afferent signals provided by sensory receptors might activate (after some low level processing) some portion of the self model data

structure, which might then activate another portion of the self model with efferent connections to low level parts of the motor control system, thereby closing the circuit on a sensori-motor response loop more quickly than the initial signals could propagate directly at the lower level from the sensory processing system to the motor activation system. This is akin to the emergence of polymodal associations, where because of physical limitations on connectivity, the pathway through a group of neurons at a high level in an hierarchical processing structure may be the *only* pathway. At this point, however, the account may ring slightly counterintuitive: isn't it the direct connections, not consciously mediated, which seem to offer the fastest response time? If so, how can a higher level self model connection between two compressed systems be both faster than the lower level connections and behaviourally efficacious?

We can clearly answer the first question in the affirmative, with human reflex reactions occurring across direct connections at the spinal cord or brain stem as the paradigm example of rapid low level responses. We shall return to this presently. The best answer to the second question (an answer which forms the basis of the final two selective advantages we will consider) is that, as we noted above, the actual physical limits on axonal and dendritic arborisation[49] mean that some of the connections which might exist between compressed representations at the self model level may not be present at all at the lower level of the systems being represented. Alternatively, they may be present, but they may be so sparse that the time required for the buildup of a front of spike activity sufficient to prompt a response in another system connected to it at a low level might exceed the time necessary to invoke a similar response through a higher level self model "short cut". This fact that the self model may help overcome the physical connectional limits at the lower levels is an important point which gets further attention in the next section.

8.4 Connections are Hard to Come By

With some 100 thousand millions (10^{11}) of neurons and about 1 quadrillion (10^{15}) synapses (Thomson 1985), the human brain has been

---

[49] The latter is analogous to what is referred to in the VLSI context as "fan-in". Interestingly, the physical fan-in limitations for both current silicon chip fabrication techniques and human neural wetware are apparently of about the same magnitude.

called "an asynchronous, nonlinear, massively parallel, feedback dynamical system of cosmological proportions" (Kosko 1992, p. 13). It is the most complex object in the humanly known cosmos. Yet, a simple calculation reveals that the average connectivity of a single neuron in the human nervous system is no more than 10,000 inputs. It is unlikely that neurons with more than an order of magnitude more connections are very common. With an actual physical limit on the number of synapses which can fit on a soma and its dendrites, there is a limit to the number of direct connections between neuronal groups which can be made at a single level. Making more connections requires moving up an organisational level so that another neuronal group representing the activity of a lower level group (or perhaps several such groups) may provide signals when the efferent connectional capacity of the lower level group (or groups) is exhausted. It is here that the architectural features of the self model come into play.

Because the representations of various models in the whole organism are lossy, the self model may maintain coherent organisational relationships between systems which would be difficult if not impossible to preserve solely at lower levels, given the physical limits to connectivity. The heterarchical nature of the self model invites the emergence of complexity which could not develop coherently in its absence. That the capacity for coherent organisation of progressively more complex relationships would endow an organism with a selective advantage is clear. But in addition to the obvious improvement in behavioural sophistication, this kind of architecture suggests another related advantage.

8.4.1 The Ups and Downs of Learning Levels

This related advantage is simply that high level complex and coherent organisation might enable a particular skill to be learnt as correlations between compressed representations at the self model level and eventually incorporated into direct, ultimately faster, connections at lower levels. For instance, a relationship between two motor actions may be grasped at a high level by the self model. The actual motor coordination of the organism might not be such that the two motor actions could be physically performed immediately with the requisite relationship (perhaps because encoding the relationship at the low level

demands the modification of a large number of connections), but the relationship could still be "imagined" at the self model level.

Over time, with activation from the self model level *down* to the lower levels, the proper motor actions could eventually be learnt by the real systems responsible for the actions. This is analogous to understanding how to do something physically and being able to imagine it instantly but still not being able to do it without practice. We can see something like this process at work in a human example of learning to perform a complex action quickly and precisely as in the study of martial arts.

8.4.2 The Ups and Downs of Martial Arts Levels

The beginning student of a martial art encounters a set of movements largely unlike any he or she has performed before. On the first few (hundred or thousand) attempts, most students cannot perform these movements with anything but a passing resemblance to what they have seen, but gradually they are able to analyse at a *high level* the sequence of physical changes which occur in the instructor's body as he or she demonstrates a technique. This high level understanding of what the student is supposed to be doing is the first step in performing the technique correctly. By "high level" in this context, we mean the student comes to understand composite *sets* of physical changes "from the outside": the left hip rotates forward, the left food slides and pivots to the outside, the right foot skims in an arc across the floor as the right hand rises up slightly from the right hip, the right foot is planted and the right hip begins to rotate forward, etc. The student needn't understand what muscle actions must be performed, say, to rotate the hips, since this can be understood as a higher level compressed representation of motor skills which the student already has developed. The student also doesn't immediately understand what entire sequences of movements would feel like "from the inside", either because not all components of the sequence have been learnt before or because they have not been experienced in the particular sequence being learnt.[50] However, grasping the sequence of

---

[50] This does not count against the idea that the self model may "imagine" being in states it has never before experienced, however, since there is substantial difference in complexity between imagining a hip twist, say, and imagining a hip twist rapidly followed by a foot slide and pivot rapidly followed by another foot slide accompanied by a small hand movement rapidly followed by a foot plant, etc.

composite movements at this high level is the first step in establishing the lower level coordinations which enable the student to mimic the technique correctly.

With this high level grasp of a particular technique, the student may begin to experiment physically with performing movements which approximate the look and feel of the sequence in place at the higher level. The student learns what approximations of the physical movements feel like "from the inside" and learns to match them with the desired appearance "from the outside". Gradually, the difference between the desired and the actual performance of the technique is decreased until, with continued high level "supervision" by the self model (more on this kind of "supervised learning" in Chapter 10), the technique may be performed in a way roughly resembling the demonstrated one.[51] But even at this stage, the training process is far from complete.

The task now is to obviate the need for this conscious high level "supervision" to ensure the match between performance and high level understanding. While the student has moved from being confused about what physical changes are taking place in the instructor's body and thus not knowing how even to begin to emulate them to understanding what the changes are and being able to imitate them with conscious direction, usually the student's performance cannot be fast enough or smooth enough while still under this supervisory control of the self model. (Taking a cue from our previous discussion, this may be due to unfeasibly large cross-level propagation delays in the heterarchical neural architecture.) The next step, then, as we have indicated, is to remove the need for this supervisory control.

This amounts simply to building directly, at the low level of the sensori-motor control loops, the appropriate connections for "automatic" performance of the muscle actions required to perform the given techniques. Because the brain does not undertake architectural changes in an "instructionist" fashion—that is, it does not determine what changes must be made and then go and change them—this process of establishing

---

[51] In practice, of course, the training process is much more complex than this, with discontinuities not only in the correspondence between the student's modelling of techniques and actual performance of them but also in the correspondence between the student's model and the actual techniques being learnt. That is, the student becomes progressively more aware of the subtleties of the instructor's technique so that, in a way, the student's actual performance of a technique is chasing a "moving target".

the direct low level connections is primarily nonconscious and results only from repeated performance of the technique to enable cellular level mechanisms to do their work in strengthening the capacity for mutual excitation between neuronal groups whose activations have become correlated through the course of that repetition. This point in training, then, can be seen in part as the development of lower level neuronal group repertoires in the sensori-motor systems appropriate for executing techniques which hitherto required the intermediation of connections at the self model level.

This may be thought of as "pushing down" into the lower levels the understanding initially made only at the self model level. The low level architecture was not suited for learning such complex movements, but the self model, capable with its lossy representations of grasping the "bigger picture" was able not only to model the technique at a high level, but it was instrumental in eventually inducing the appropriate complex connections at the lower level.

Rather than "pushing down" the modelling of the technique from the higher level, this process might alternatively amount to "raising up" the level of the self model. The high level pathways may with practice become dedicated to providing the signals which cannot be sent directly at the lower levels, and their connections with the rest of the self model may largely atrophy, requiring that whatever previous functional rôle they served at that level be taken over by other neuronal groups. Here, degeneracy and variance in the group "recruited" to serve as the high level shortcut may enable selective mechanisms (pace Edelman, see chapter following) to "choose" a new group, largely already appropriately configured, for the high level rôle. While part of the self's instantiating wetware, then, may be lost to the low level task during this kind of training, other groups are ready to take its place, and little of the self model's information processing capacity need be sacrificed for any significant length of time.

The student has now given the self model new representations and the system new coordinations with which to work. Just as coordinations acquired at a low level induce representations at the self model level, new understandings at the self model level may induce the development of coordinations at the lower level. This process in the martial arts example endows the experienced student with the ability to analyse higher level

combinations of techniques and to understand the tactical implications of this or that combination as compared to another. This is akin to the development of a new vocabulary, and, ironically, the process of analysing and mimicking and transcending conscious supervision in the advanced student exactly mirrors the process in the complete neophyte, except that the complexity of the objects of analysis differs in the two cases.

8.4.3 Learning Quickness and Quickness in Learning

Thus, the improvement to learning enabled by the self model bears directly on the question we raised a short time ago (namely, aren't the low level connections the fastest ones?). When there is sufficient low level connectivity to begin with, the low level direct connections are fastest because they may result in faster behavioural response than responses mediated by the self model. When this low level connectivity is lacking, where the low level connectivity is such that responses invoked solely through low level activity would be either very slow or nonexistent, the self model can improve response time by offering high level "short cuts" to appropriate responses. And finally, as exemplified by the martial arts example, the self model can help the lower level to learn the desired connections and ultimately improve the response time further by handing the processing back over to that lower level. This last process can be interpreted either as lowering the level of the processing or as raising the level of the self model; in the martial arts example, this allows the martial artist to process more complex behaviours because the learnt techniques are available as elements in an expanding "vocabulary" of representations in the self model.

Of course, while learning martial arts may improve a human's chances of surviving and reproducing, such complex endeavours are hardly relevant to the course of phylogenetic development through the history of biological life on Earth! But while the example focuses on the rôle of the self model in acquiring some very complex skills, it differs from more selectively relevant examples only by degree. By the same process we've explored in the martial arts example, the self model may aide the learning of any complex skill which the lower level neural architecture is not immediately suited to grasping. As we know from both classical and operant conditioning, complex skills are best taught as composites of simpler skills learnt first; simple low level sensori-motor systems are not

suited to recognising and emulating novel and highly complex patterns, at least in part because of the limits to neural connectivity we have noted previously. Self models, however, are better at this task and may enable the emulation of a pattern, perhaps (when relevant representations are at a suitably high level of an hierarchy) after even only one presentation of an example. In the end, they may enable the integration into the behavioural repertoire of lower level coordinations which could be got only very slowly—if at all—by direct learning at the lower level. Thus, the self model is a valuable tool for an organism which may need to copy the behaviour of another organism (such as a parent) or which may benefit from developing complex responses of its own after observation of some environmental scenario. Most higher organisms arguably fall into these categories and could thus benefit from exploiting the capabilities of a self model.

Having explored these various selective advantages of self models and thus, in the absence of any immediately obvious arguments against them in a biological context, their *prima facie* plausibility in terms of evolutionary biology, we will soon turn to analysing the particular architectural methods by which self models may be implemented. By way of transition, we will first make some notes about the cognitive neuroscientific account of perceptual categorisation of Gerald Edelman; in many ways, his theory is at the halfway point between low level neurology and higher level cognitive science, and our self model theory may be seen as both a higher level extension of Edelman's work and a philosophically satisfying bridge of this gap. We then discuss developments in artificial neural networks and how they may help us understand cognition in real neural systems, together with some comment on the debate between proponents of classical symbolic paradigms and proponents of strictly distributed systems. Finally we move on to explore some simple examples of the actual architectural layout of various components of the self model.

9

Perception and Neural Darwinism

Before getting on to the architecture of self model implementations, it is useful to look at a similar kind of theory intended as an account of perceptual categorisation. Our own view of self models can be seen in a way as a higher level extension of the perceptual categorisation theory both in the sense that self models rely upon architectures being in place to subserve their perceptual categorisation and in the sense that self models are intended to bring together the low level perceptual categorisation and the higher level conscious experience of *doing* perceptual categorisation. Also, many of the same neurological inspirations motivate both theories. Finally, just as the physicist studying the weak nuclear force may benefit from gaining an understanding of electromagnetism and the kinds of questions answered by the theory of electromagnetism—perhaps with the goal of eventually incorporating them each into one larger theory—so too may we benefit from understanding perceptual categorisation and the questions which the theory helps to answer. Getting some kind of grasp on the established theory will help us better to understand some of the ways our own self model theory should be formulated and the kinds of architectural questions and features it must accommodate. Our purpose, however, isn't to explore the theory in all its complexity, but merely to note some of the cytoarchitectonic features on which it relies and to see both its compatibilities with existing empirical data and its relationships to our own broader project.

9.1 Neural Darwinism

Arguably the richest cognitive neuroscientific account of perceptual categorisation in the human brain is the theory of neuronal group selection—so-called "neural Darwinism"—advanced by Nobel laureate Gerald M. Edelman and colleagues. (Edelman 1978, 1981, 1989a, 1989b;

Edelman and Reeke 1982; Edelman and Finkel 1984) One of the most important tasks of the nervous system is to learn to categorise the rich landscape of perceptions in a world devoid of pre-existing "labels". Many of an organism's most basic actions require the ability to differentiate object from background, friend from foe, beneficial from harmful. An organism must be competent to recognise two different presentations of the same stimulus and ultimately to recognise similarities between related stimuli and to direct its behaviour toward them appropriately. Edelman offers an account, in terms of organisational features of large populations of neurons, of the various mechanisms which contribute to these extraordinary abilities. The theory of neuronal group selection is especially concerned to explain the kind of polymorphous (Ryle 1949) categorisation considered by Wittgenstein (1953; see also Pitcher 1968) with respect to games, where membership in a category is allowed when any m out of $n \geq m$ possible disjunctive properties are displayed.[52]

Edelman's theory doesn't answer the questions about sensation which the self model view is meant to illuminate, but it is useful to notice features of Edelman's account which bear on the self model's architectural features that we will shortly explore. In particular, it is useful to keep general properties of the neuronal group selection theory in mind when we finally pin down the self model architecture in order to monitor the compatibility between the two approaches.

9.2 Socialist Neuroscience?

The theory of neuronal group selection is an application to neuroscience of the kinds of population ideas (see Mayr 1982 for overview) which we discussed in the previous chapter with respect to the evolutionary plausibility of self models. The most significant selectionist feature of the theory is that selection operates primarily on *groups* of hundreds to thousands of closely interconnected and functionally related neurons. In contrast to the work of Changeaux and colleagues (Changeux and Danchin 1976; Changeux, et al. 1984), whose selectionist model of development and epigenesis works on the basis of eliminative selection of

---

[52] Apparently humans generally employ much more complex strategies in categorisation than the rigid lists of individually necessary and jointly sufficient conditions of standard set theory. (Rosch and Lloyd 1978, Smith and Medin 1981)

individual neurons, the theory describes mechanisms for implementing both positive and negative selection of entire neuronal populations.

Edelman's theory is unique in this emphasis on entire neuronal populations and the features which he builds into them and into the interactions between them. Perhaps the two most important features are degeneracy and the kind of phasic reentrant signalling which we discussed above for cross modal associations in our own self models.

Degeneracy is the availability in each repertoire of functionally related neuronal groups of an abundance of isofunctional but not isomorphic structural variants. That is, in each large "grouping of groupings" of neurons, there are many groups which respond very similarly but which exhibit different architectures. This structural diversity originates epigenetically in prenatal development and appears in part due to various selective pressures regulated by cell and substrate adhesion molecules and their effects on cell division and death, movement, and differentiation. Edelman is careful to distinguish this notion of degeneracy from the related concept of *redundancy*, which describes groups of structures which are both isofunctional *and* isomorphic. Both degeneracy and redundancy offer reliability and consistency in systems made up of stochastically variable or unreliable componentry. (von Neumann 1956; Winograd and Cowan 1963) Perhaps more importantly in our context, degeneracy evinces the kind of variation present in any natural population owing its constitution to selective pressures, and it offers a wider range of possible responses to completely novel stimuli while preventing the kind of "overfitting" which may occur when an architecture adapts inflexibly to a given set of requirements.

The most significant feature of interaction between these neuronal groups emerges postnatally with epigenetic modifications to the synaptic connectivity between (as well as within) neuronal groups. Various mechanisms—more on synaptic changes presently—contribute to the formation of reciprocal reentrant connections between receptor sheets for different modalities (and capable of what Edelman dubs independent "disjunctive sampling"), motor ensembles, and so forth which are correlated in their output activity. Edelman believes reentrant connections emerge in response to selective pressures at the group and cellular levels and facilitate the kinds of phasic signalling which he

indicates subserve the maintenance of coherent spatiotemporal relationships between environmental input and the organism's responses.

Edelman argues convincingly that such a model of neuronal group selection is a biologically plausible route to the emergence and maintenance of sophisticated networks capable of polymodal perceptual categorisation in an unlabelled world. His account of selective pressures operating on degenerate groups and the development of reentrant signalling between groups responding to afferent signals from different modalities is vastly preferable to instructionist or computationalist alternatives which show little promise of being accommodated by the neuroscientific data. In this sense, Edelman's approach stands out as the most fruitful in a limited field of theories which attempt to tackle these kinds of problems.

In addition to the emphasis on features of group selection, Edelman's theory is also unique in its treatment of individual changes at the neuronal level, and it is both compatible with and partly inspired by a large body of cellular studies (for overviews, see Edelman, et al. 1985; Purves and Lichtman 1985). In contrast to the somewhat primitive cell assembly theory due to Hebb (1949, 1980, 1982; see also Hayek 1952)—which, incidentally, is the basis for almost all development of artificial neural networks which exhibit plasticity and which moreover cannot be entirely correct (see Wigstrom, et al 1982 for hippocampus data)—the theory incorporates differential pre- and post-synaptic rules for updates to synaptic efficacy. (Finkel and Edelman 1985) It can accommodate both homosynaptic change (i.e., change in plasticity as a result of activity at the synapse concerned) and heterosynaptic change (i.e., change in plasticity as a result of activity at nearby synapses on the same neuron). There is good experimental evidence for both types of modifications at the ultrastructural and biochemical levels (Fifková and van Harreveld 1977; Desmond and Levy 1981; Vrensen and Nunes-Cardozo 1981). Finally, Edelman's mechanisms are incompatible with the simple chemoaffinity patterning models credited to Sperry (1963, 1965).

9.3 Beyond the Group—Post-Socialism?

In the end, the theory of neuronal group selection is successful in many ways as a first step towards answering some of the most puzzling

questions about perceptual categorisation. In just one example, Edelman has described a simple mapping network operating only with a postsynaptic rule, the behaviour of which mimics very well several experiments reported in the literature. (Merzenich, et al 1984) The self organising network apparently emulates roughly the emergent organisational characteristics of area 3b of the somatosensory cortex of the squirrel monkey, an area dedicated to mapping of the glabrous and dorsal surfaces of the hand. (Merzenich, et al 1983a, 1983b) Edelman is also keen to point out (1989a, p. 289) that the "classification couple" (a simple pair of feature maps with reentrant signalling) in the rudimentary Darwin II network overcomes the artificial neural network limitations famously described by Minsky and Papert (1969), although in all fairness there now exists a broad selection of other artificial neural networks available in the literature which also overcome the early perceptron limitations[53]—a broad selection which, much to the consternation of his critics, Edelman typically fails to mention in his own writing.

While neural Darwinism is a good first step towards understanding one aspect of the brain as a distributed system (Mountcastle 1978), however, many questions remain unanswered. For instance, although Edelman's theory incorporates the rôle of cortical columnar bundles in primary sensory receiving areas for mapping multidimensional properties of unimodal information to a two dimensional sheet (Hubel and Wiesel 1977), such mapping alone is apparently not sufficient to account for the relevant properties of perception (Uttal 1978, 1981). Edelman (1989a, p. 109) agrees with Zeki (1981) that further reentrant mapping with multidimensional characteristics "seems to be required". The extent of the additional requirement, however, is not quantified. The impression we get from Edelman's presentation is that most of the ideas included in the theory are on the right track and that most of the organisational and cytoarchitectonic features it predicts will probably turn out to be confirmed by further experiment, yet there are still significant elements of the "bigger picture" missing. I believe Edelman is correct in singling out reentrant signalling between classification couples and the various selective mechanisms which help to create them as some of the most important underpinnings of any complete account of perceptual categorisation, but it

---

53 These limitations served nearly to kill the field of artificial neural networks altogether. Fortunately over the last decade or so the field has grown robustly.

is disheartening to see, despite the technical impressiveness of Edelman's achievement, just how much work remains to be done.

More importantly for our project, Edelman's theory does not begin to explain how it is that it *feels* the way it does to be a subject of sensation engaged in the kind of perceptual categorisation which the theory seeks to explain. After all, basic categorisation can be performed by the simple kinds of artificial self organising feature maps (SOFMs) to which we turn in the next chapter, yet there is no reason to believe that such data structures experience sensations of any kind—and plenty of reasons to believe they do not. Edelman's theory fills the important rôle of helping us to understand the perceptual abilities of humans, but it doesn't explain the relationship between perceptual categorisation and the sensation of being an individual in a state of *doing* perceptual categorisation. This is where our own self model approach can serve as an extension of the neuronal group selection framework, a means for understanding what *else* there must be, in addition to mechanisms for enabling perceptual categorisation, in order that such categorisation be accompanied by sensation. The self model theory we have been exploring complements Edelman's theory nicely, and hopefully it may contribute to the development of more sophisticated psychological models along the lines of those based on the Edelman-style evolutionary selective view of the brain. (Rosenfield 1988, Goertzel 1993)

Next we will flesh out our high level description of the self model approach with explorations of their lower level architectural features. We begin with comments on the artificial neural network framework in the terms of which much of the rest of our discussions will be couched.

10

Simple Neurons Doing Simple Things

We shall try to avoid allowing the present chapter to degenerate into a tutorial on connectionism and artificial neural network development, but it will be useful to pursue a few points, both technical and philosophical, which will be directly relevant to our impending foray into self model implementation. Seeing self models through the eyes of artificial neural networks will help us to pin down the essential functional relationships of the neural architectures without being too distracted by real biological factors which are poorly understood and which may not be entirely necessary for implementing the kinds of data structures which we are exploring. Having said that, understanding the artificial neural network framework will also reveal something of the extent to which we may be missing out on important factors of biological reality in couching our discussions *too* much in connectionist terms.

For the interested reader, there is already a growing body of literature about artificial neural networks and philosophical questions of mind (Clark 1989; Horgan and Tienson 1991; Ramsey, et al 1991; Clark and Lutz 1992) and about enlightening philosophy with a cognitive neuroscientific approach (Churchland 1986), as well as about broader questions of philosophy and artificial intelligence approaches (Boden 1990) and technical matters which are combinations of cognitive neuroscience and philosophy of mind (Churchland and Sejnowski 1992) or cognitive science and neurobiology. (Gardner 1993)

10.1 Artificial Neural Networks—The Tutorial

The basic approach in the development of artificial neural networks is to reduce the behaviour of a neuron to that of a more or less functionally simple input/output node, with outputs usually in the range

[0...1] or [-1...1], and then to connect these nodes together in networks, usually simulated in software on a digital computer, across which nodes can exchange signals. Sometimes these networks are hardwired in VLSI chips, and some effort is even being made to develop more biologically plausible models of single neurons for implementation in silicon. (Mahowald and Douglas 1991) The point of artificial neural network development is that when these nodes are connected together in appropriate ways, particular nodes may be excited with some kind of input, and the network may process the input in such a way as to give a useful output on some other set of nodes.

10.1.1 Connections are the Spice of Life

These idealised nodes communicate with each other by means of weighted connections representing synapses. Weights indicate the synaptic efficacy of a connection between two nodes and are multiplied by the numerical output of the first node to determine that node's contribution to the activity of the second node.[54] To calculate the output of any individual node, all its weighted inputs are first summed. An output function is then applied to this total weighted input to determine the activation level of that node and the value which it will provide through weighted connections to subsequent nodes. The activation level may be described by a simple threshold function, where the node gives maximum output when its weighted inputs exceed a certain threshold value and minimum output when they fall below this level. Instead, it might be a simple straight line output, where output rises linearly with input, a nonlinear sigmoid function, or some other variety. Individual nodes may also have a certain bias—corresponding to a real neuron's base firing frequency—an initial proclivity to fire or to remain at rest, which may be implemented either as a single value associated with each node and which is added to the weighted sum of inputs each time output is calculated or, equivalently, as a weighted afferent connection to a special node which provides a continuous maximum output.

---

[54] Weights may be either positive, representing excitatory connections, or negative, representing inhibitory connections. Likewise, weighted inputs may be either positive, indicating an excitatory postsynaptic potential (EPSP), or negative, indicating an inhibitory postsynaptic potential (IPSP).

10.1.2 Learning is the Spice of Connections

If this were the whole artificial neural network story, networks would really only be able to push values around, and designing useful ones would be a very complex process demanding careful analysis to determine what arrangements of connections and weights would give the desired output. What makes these neural networks interesting is that by incorporating some rule for updating the weights of connections between nodes as the network is presented with different kinds of inputs, a network may *learn* and adapt to give more useful output on future presentations of similar input. Learning rules may be either supervised, in which case the network is "told" what outputs are appropriate for particular inputs and the network is responsible for altering its weights to match the inputs to the outputs, or they may be unsupervised, in which case the network has no information about what outputs should correlate with what inputs and simply organises itself in a way which is more or less coherent according to the character of the learning rule. These simple learning rules are analogous to the mechanisms which effect arborisation in a network of real neurons.

On first blush, it should be only networks with unsupervised learning rules which will be of interest to us, concerned as we are not so much with what kinds of interesting computational feats may be accomplished with artificial networks but instead with what kinds of data structures may be implemented by real biological systems in which there is presumably no "supervisor". (As we noted in the prior discussion about neural Darwinism, categorisation must take place in an unlabelled world; providing supervision for a network in the form of desired outputs is tantamount to offering it such labels.) Later we will see that there is a way in which this is not entirely true, but for the moment we will examine an example network which may serve to illustrate the general idea of self organising networks.

10.2 Learning With No Exams

The following network is an example both of unsupervised adaptation and of the trend in artificial neural network development to create networks endowed with more biological plausibility. The network is similar to others developed with the Kohonen (1984) algorithm which

self organise to produce coherent feature maps of the input data. I have reprinted the following from Mulhauser (1993b),[55] with some corrections and minor modifications to enhance continuity with the discussion we have made so far.

10.2.1 Self Organisation and Biological Plausibility

As we have previously noted, artificial neural networks are inspired by the networks of nerve cells in brains but are designed for use on electronic computers. Unfortunately, in the quest to simplify artificial neural models and reduce the computational overhead required for their execution, a number of poorly understood characteristics of biological neural networks have been abstracted away or ignored altogether. The model I propose showcases novel approaches for implementing certain of these neuronal characteristics which may return performance improvements when integrated into existing artificial neural networks. The model makes minimal computational demands and is unique in combining lateral connections within a layer to represent gap junctions and ephaptic interactions, a fatigue factor for each node, and a Hebbian (1949, 1980, 1982) learning rule.

While I am not yet able to report results of empirical testing of the network architecture and learning algorithm I describe[56], structural similarities to other networks suggest that we might expect the net to produce population coding which resembles something like a cross between Kohonen's modified competitive learning strategy (1984) and principal-components algorithms. By 'population coding', we mean that activity in response to a particular input is concentrated in a particular local group of neurons. Such coding may offer computationally cheap feature extraction because the location of the active population of neurons can be specified simply.

10.2.2 Architectural Preliminaries

The network consists of an input layer completely connected to a second layer that includes lateral connections. That is, every node in the

---

55 Please see the Appendix for reprint information.

56 Initial programming and testing of the network was undertaken near the end of the 1992-93 academic year by the Computing Studies Department of Napier University (Edinburgh, Scotland), but the programme was not continued in the new year, and I cannot report any substantial experimental data.

input layer is connected to every node in the second layer, and nodes in the second layer are partly connected to each other. It is within this layer that we should expect population coding to emerge. Connection weights are in the range [-1...1] and node outputs are in the range [0...1]. Node outputs, except for the fatigue modification which I describe below, may be calculated from the linear sum of weighted inputs according to a sigmoid curve or other standard function. Note that no derivatives are taken of the output function (as some learning rules demand). Thus it may be desirable to choose an output function which allows more information effectively to be represented in a node's output value, much as information may be represented in the output frequency of a real neuron.[57] This might even provide a few of the benefits of pulse coded networks. (Gluck, et al 1989; Duchateau and Lansner 1991; Gustafsson, et al 1992; Kosko 1992) Biases are not addressed but may be implemented in standard ways if desired.

10.2.3 Footballs and Pyramids

In biological neural networks, local neuronal activity is partially homogenised by means of gap junctions and ephaptic interactions. The former are actual physical connections between adjacent cells made by large macromolecules which extend through both cell membranes and contain water-filled pores. (Kuffler, et al 1984; McCormick 1990) Ephaptic interactions do not require a physical connection between nerve cells, but they have a similar effect: the electrical currents set up by the flow of ions across the membrane of one neuron may induce electrical currents in nearby cells. (Hille 1984; Kuffler, et al 1984)

While the strategy used by Kohonen to enable physically adjacent nodes in a competitive learning network to code similar input patterns is one possible abstraction of this feature, I propose a simpler abstraction in which the input layer is completely connected to a single layer, within which each node is directly connected to its three nearest neighbours. The architecture may be viewed geometrically as a regular pattern of hexagons,

---

[57] Because of the way lateral connections are treated in this network, smoother output functions are favoured over step functions or very steeply nonlinear ones.

as in Figure 1.[58] For simplicity in implementing the connections as a data structure, the pattern can simply be stretched as shown in Figure 2.

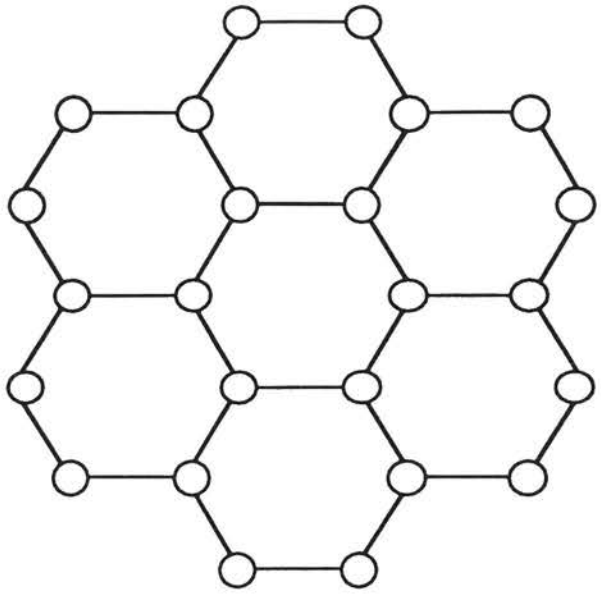


Fig. 1. Horizontal connections in the second layer

For some applications, it may be desirable to maintain the pattern of three lateral connections per node by warping the structure into a third dimension and connecting nodes shown here with only two connections to the corresponding nodes on the opposite side. The uniformity of connections might also be achieved by arranging nodes in patterns of both hexagons and pentagons and transforming the whole structure into a three dimensional football shape.

[58] Interestingly, the decision to limit connections to the three nearest neighbours was based on biological considerations and intuitions about limiting the influence of the lateral connections. One year later, Der and Herrmann (1994) confirmed in their Voronoï tesellation study that stability concerns related to the influence of lateral connections make hexagonal arrangements preferable to quadrangular arrangements for many applications.

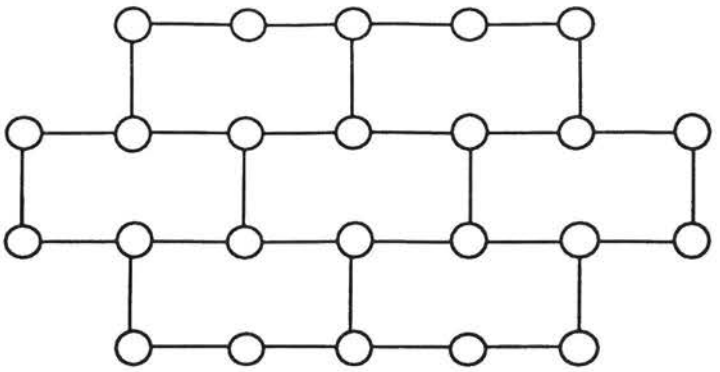


Fig. 2. Simplified horizontal connections

In implementing this connection pattern, we may treat lateral synapses as "virtual connections" and view all connections as being *between* layers. This is accomplished by introducing a third layer with a number of nodes equal to that in the layer in which we are implementing "virtual connections". We may then dispense with horizontal connections between nodes and calculate their effects in the subsequent layer. Each node is connected to "itself" in the next layer by a connection whose strength is permanently set to one. (Obviously, if biases are implemented, they must be dropped for the "copy" in this third layer; otherwise the biases provided to the two nodes representing the same laterally influenced node would be additive, and this convenient computational trick would not work.) This node also receives inputs from each of its neighbours in the preceding layer, as in Figure 3.

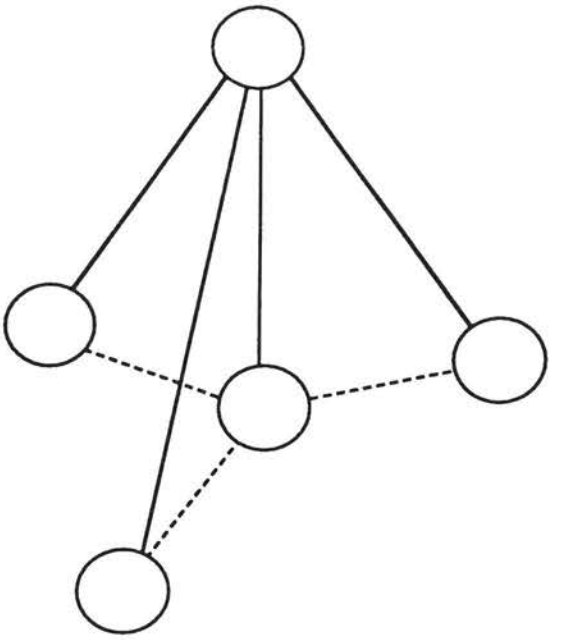


Fig. 3. Implementing virtual horizontal connections with an extra layer

We may thus calculate first the response of each node to the firing patterns of the input layer and then calculate separately the influence on it of the activity of the nodes in its horizontally local neighbourhood. There may be different advantages to setting each of the virtual lateral connections to a uniform initial strength such as 1/3 or simply to randomising them across the net.

10.2.4 Unsupervised Learning

Because of its rôle in the function for updating connection strengths, I describe first the fatigue factor associated with each neuron. Fatigue in real neurons, known as spike frequency adaptation, is the tendency of some kinds of neurons to decrease their firing frequency in response to sustained depolarisations. This behaviour may prevent the same neurons from becoming active in representations of too many distinct input patterns. For the present model, I suggest a fatigue value ϕ_x in the range [0...1] (where larger numbers indicate more fatigue) which is applied to calculate a "real output" O_x for a node x by a straightforward modification of an output out_x calculated with a sigmoid or other function from the weighted sum of inputs to the node:

$$O_x = \text{out}_x(1 - \phi_x) \tag{1}$$

When calculating a new fatigue value ϕ'_x for a node x, we should like to take into account its recent firing history as represented by the previous fatigue value as well as the magnitude of its present output. If its present output is low relative to its fatigue value, the fatigue should fade quickly. Likewise, if its present output is high, fatigue should achieve significance rapidly. This can be modelled as in equations (2) and (3),

$$\phi'_x = \phi_x + \varepsilon(O_x - \phi_x)^2 \qquad \text{when } O_x > \phi_x \tag{2}$$

$$\phi'_x = (\phi_x)^2 \qquad \text{when } O_x \leq \phi_x \tag{3}$$

where ε is a parameter set in advance to adjust how significant an influence fatigue is allowed to become. Fatigue values should be set to 0 before the network is trained.

The learning rule I describe is inspired by Hebb's theory that synaptic coupling increases when the activity of converging network elements is coincident.[59] We would like to increase the weight of a connection between two nodes when the nodes produce similar output. A high correlation paired with a low present connection strength should correspond to a large increase in weight, while a low correlation paired with a high connection strength should correspond to a large decrease in connection strength. We seek a relationship something like Figure 4, where the x-axis represents the present connection weight, the y-axis the correlation between the nodes' outputs, and the z-axis the amount by which the connection should be updated.

---

[59] Recall that the Hebb rule cannot be entirely correct. Curiously, however, something closely related to Hebb's theory is the basis of nearly all artificial learning networks today which claim biological plausibility.

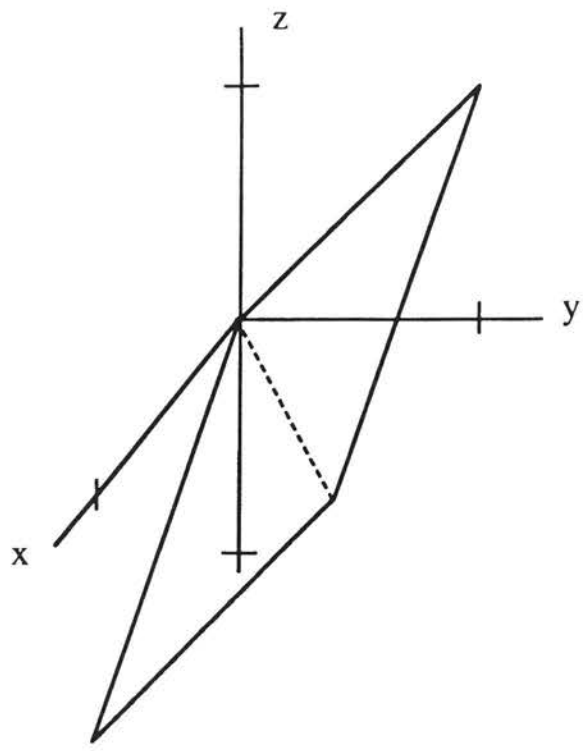


Fig. 4. Weight updating

The relationship might be nonlinearised as in Figure 5 to increase the responsiveness of the network, although such a nonlinearisation would need to be applied carefully to avoid undue influences on the behaviour of lateral connections and to avoid increasing the sensitivity of the network to a level where it becomes unable to converge on a stable pattern of connection strengths.[60]

[60] I believe it was Joos Vandewalle of the Katholieke Universiteit Leuven who first noted at the original presentation of this network that it is not guaranteed to converge. As I indicated at the time, however, I am concerned more with biological plausibility than computational utility, and there is no guarantee that biological networks converge.

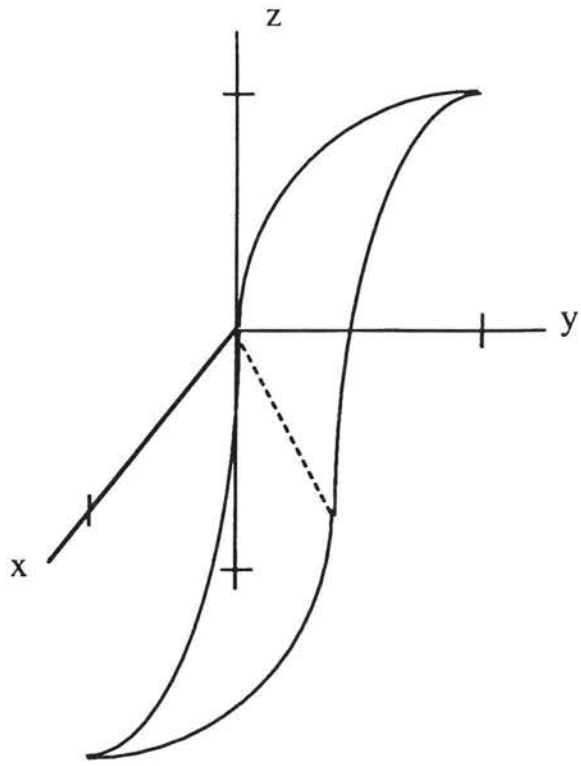


Fig. 5. Nonlinearised weight updating

I suggest that equation (4) establishes the relationship indicated in Figure 4 between a $w_{x,y}$ scaled to [0...1] (x-axis), representing the present connection weight between nodes x and y, $1 - |O_x - O_y|$ (y-axis), representing the correlation between the outputs O of nodes x and y, and $\delta(w_{x,y})$ (z-axis), the amount by which to update the connection strength.

$$\delta(w_{x,y}) = \sigma\left[1 - \left|O_x - O_y\right| - \frac{w_{x,y} + 1}{2}\right] \tag{4}$$

Here $w_{x,y}$, which ranges over [-1...1], is scaled to the range [0...1] by the last term inside the brackets.

The σ term relates to the implementation of the fatigue factor. We are seeking a strategy which will avoid large modifications to a connection weight when both nodes in question are highly fatigued. This might occur, for instance, in situations where a similar input pattern has been presented over and over again and the nodes coding it have begun to decrease their outputs in response to fatigue. Equation (5) indicates one method of implementing the strategy:

$$\sigma = [1 - \min(\phi_x, \phi_y)][\max(O_x, O_y)] \quad (5)$$

The second bracketed term in (5) also ties the degree to which a weight is altered to the strength of the nodes' outputs. Thus, high correlation must be coupled with high output magnitude to invoke a maximal change in connection strength.

Finally, returning to equation (4), for applications where we are concerned with an eventual settling down of the network's connection weight activity—with or without convergence—we may add the quantity in (6) to the front of the bracketed term in order to decrease gradually over time the degree to which weight changes are made. (This results in something akin to psychologists' primacy effect but not the corresponding recency effect.) Here, η is the total number of inputs to be presented and ι is the the index of the current input.

$$\frac{\eta - \iota + 1}{\eta} \quad (6)$$

10.2.5 What's Wrong With This Picture?

While the techniques described here are inspired by theories of how real living neurons function, they are not without their limitations, in part because these theories may not accurately reflect real neural behaviour. The onset and decay of fatigue in real neurons is poorly understood, and the effects of ephaptic interactions and gap junctions are difficult to quantify. The characteristics of dendritic growth are actively researched, but it is impossible to say yet just how badly flawed Hebb's theory of synaptic change may be.

The network implementation makes minimal demands on computational resources and may lend itself to the construction of larger but faster nets, yet the specific strategies at work in the net might prove problematic. For instance, the network may prove too sensitive to the choice of an ε value in equation (2). Equations (2) and (3) may allow fatigue to become significant too rapidly or to decay too slowly. Incorporating data from single neuron spike frequency recording could help fine tune these parameters. Also, depending on the types of data presented to the input layer, it may be necessary to place a bound on the

value of the virtual horizontal connection weights to prevent interactions within this layer from becoming too significant. Finally, the virtual horizontal connections may produce nonstandard relationships between the number of nodes in the input layer and the desirable total number of connections in the network.

This architecture and learning algorithm invite further research into the performance of the network I have defined as well as the performance of other networks modified with some of the techniques I have described. Virtual horizontal connections as a simple means of encouraging population coding should be tested and compared with other established methods, and the usefulness of fatigue factors to promote uniform distributions of activity may be examined in any net where population coding is the main aim. The learning algorithm itself may be tested and improved as a substitute for less biologically plausible algorithms currently enjoying widespread use.

Other promising avenues for research with this type of network include the implementation of mechanisms for growing new horizontal connections (perhaps with a speed inversely related to the distance between the two neurons in question) to allow synchronisation and lateral reinforcement between spatially separated but functionally related network nodes, interlevel feedback, and the implementation of recurrent feedback of the output activity back to the input layer.

10.3 The Tutorial—Part Two

We can see from this brief detour into artificial self organising neural networks that the mathematical abstractions can be rather far from biological reality. Nonetheless, as we noted previously, they can help us get at the essential functional processes which are going on in complex neural systems while sheltering us from a portion of the low level cytoarchitectonic hustle and bustle. (Of course, we must always remember that a good deal of this hustle and bustle is undoubtedly functionally relevant, and we must keep returning to biological reality as we seek to develop networks which reflect to increasingly greater degrees the capacities of real nervous systems.) Recently, another sort of neural network development has emerged as a serious contender in terms of both biological plausibility and computational ability.

Rather than using a learning rule to alter connection strengths in a network of fixed architecture, these networks are developed by applying an artificial version of the evolutionary framework we have discussed previously to generate populations of fixed plasticity networks suited for a particular task. Typically a bit string "genome" represents the architectural layout and connection strengths of a network, and a fitness function determines how effectively a generated network performs the desired task. Initially, the genomes for a population of networks are generated randomly, and the fitness function is applied to select some percentage of them which are most up to the task. Usually, of course, the randomly generated initial networks are not very good at doing anything at all! But these selected genomes then serve as the "stock" for generating a new population of genotypes: through a genetic recombination algorithm, sections of successful genomes are combined in a way akin to real biological crossover, with a generally very small amount of mutation (usually random flipping of bits in the genome string). The fitness function is applied again to select the new networks best suited to the task, and once more the genetic material of these networks serves as the basis for generating a new population. After some generations—often surprisingly few—highly capable networks may emerge. While this is hardly a plausible rendition of real evolution, the technique has proven itself extremely powerful in automatically generating populations of networks for handling particular problems.

In what follows, we see an attempt at making such neural network development slightly more plausible, and the basis becomes clear for the side comment above that "supervised" networks are not altogether useless for understanding real biological systems. The emergence of hybrid architectures as described below is also helpful for understanding the biological plausibility of the self model architectures we will soon explore. This kind of pseudo-supervision by a classical subnetwork is also an excellent example of the instantiation of the feedback which we noted earlier could be so useful to the self model. Following the style of the previous network illustration, I have reprinted the following from Mulhauser (1994b),[61] again with some corrections and minor modifications to enhance continuity with the discussion we have made so far.

---

[61] Please see the Appendix for reprint information.

10.4 The Evolutionary Goal

Present development of artificial neural networks, whether by genetic algorithms or by traditional learning methods, is grossly inadequate as a picture of biological reality. Network design with genetic algorithms ignores the rôle of ontogenetic behaviour adaptation in response to the characteristics of the particular environment in which an individual phenotype finds itself. Genetic algorithms typically are applied to generate zero plasticity networks incapable of ontogenetic development. (This is the case even for the so-called "growing networks" of Nolfi and Parisi 1992 in which the ontogeny of a network is directly, albeit nonlinearly—see also Langton 1992—encoded in the genotype and is independent of environmental factors unique to particular individuals.) Conversely, network development with traditional learning methods generally ignores the fact that real biological organisms learn within boundary conditions set by the organism's genotype and phylogenetically adapted by recombination, mutation, sexual selection, and environmental pressures. The boundary conditions of traditional learning networks are set not by the power of evolution but by human designers taking "educated guesses" at appropriate architectures, learning algorithms, and connection patterns.

But it scarcely needs pointing out that real adult phenotypes are the product of both ontogenetic *and* phylogenetic development. The living organisms we encounter every day are genetically endowed at birth with a wealth of characteristics which evolution has determined are beneficial for their survival and reproduction, but within the bounds laid down by their genotypes they are also capable of adapting to changes in their environment. They are capable of *learning*. Thus, insofar as the theoretical frameworks offered by either the genetic algorithms paradigm or the traditional learning paradigm claim to be biologically plausible, they are incomplete, and insofar as they claim to be complete, they are not biologically plausible.

10.4.1 Paying the Evolutionary Piper

I suggest a broader theoretical framework within which artificial neural network design mimics the natural features of both genetic coding

and environmentally induced adaptation by individuals. On this view, genetic algorithms should be applied to genotypes which code not only for the standard parameters describing nodes, thresholds, and connections (or simple dendritic growth for growing networks), but also for characteristics of dendritic plasticity. These characteristics might include the learning rate of a globally defined Hebbian rule as well as a measure of the capacity for growing new connections. The fitness function may then be applied to a *phenotype* grown nontrivially during a "learning phase" in a dynamic environment. It might give preference to phenotypes which adapt smoothly to environmental changes, such as a motor control network which could not only manoeuvre a robot arm toward a given point but could also adapt to avoid obstacles introduced into its path.

From artificial neural nets developed within this theoretical framework, I believe we may gain some insight into the emergence of classical symbolic processing in real intelligent organisms. I propose that evolution creates hybrid architectures of high and low plasticity connections in which arrays of neurons with low plasticity connections might embed very simple classical processors such as basic logical connectives. This view takes theoretical support from the idea that low plasticity subnetworks implementing classical functions may be the most efficient "building blocks" on which higher plasticity distributed hybrid networks could rely. Given that genotypes specify at least some, if not all, characteristics of dendritic growth (by virtue of specifying the structures of cells themselves), I believe it is highly *implausible* that Nature could have failed to exploit this elegant way of mixing the best attributes of connectionist and (neurally implemented) classical systems.

This is not to say that genetic algorithms may *only* generate networks exploiting classical processes, for this is clearly not the case. If adult organisms never had to learn by experience, never had to remember information or respond to situations radically different from those which influenced the phylogenetic development of their predecessors, genetic algorithms might have provided for the entire repertoire of behaviour of adult organisms with nonlearning networks operating with any balance of obviously symbolic or distributed principles. But in the real world higher organisms are not entirely hardwired by their genotypes. I suggest simply that real biological development yields networks of mixed plasticity and

that in those portions of phenotypes which *are* hardwired, we may find embedded symbolic functions.

By way of example of what might be accomplished by deliberately mixing high and low plasticity connections, below I describe a motor control problem together with a manually designed speculative prototype network meant to illustrate some architectural principles which might be exploited by the automated development theory I have outlined above. It is certainly inferior to what could be generated by such a strategy, so I include it not as a solution to any particular motor control problems but merely as an example of a first step.

10.4.2 A Sample Problem

For the moment we are concerned with the problem of visually guiding some mechanical device to an arbitrary point in space. This amounts to combining information about the present visual image with information about the desired image in order to activate a motor system. Here we assume that the relationship between a given activation of the motor system and its influence on the visual image is initially unknown but could be described by a (hopefully simple) computable function. For this example, we also assume that visual information has been pre-processed in such a way that the control system is presented with an indication of, for instance, the present real coordinates of the device together with its desired coordinates.[62] The dimensionality of the problem might be increased by also including the present and desired coordinates of more than one coupled part of the mechanical device, such as both the target end and the elbow joint. The details of the specific device under motor control do not concern us. Instead, I would like to paint in broad strokes a picture of one possible neural architecture for performing this type of control task.

10.4.3 Another Architecture

The network I propose works on the hypothesis that an unsupervised Hebbian network provided with feedback about its level of success at performing a particular task might approximate the capabilities of a supervised network learning to perform a similar task. This principle

---

[62] This pre-processing does not amount to "cheating": it is simply a straightforward job for another network not considered here.

is inspired by Rumelhart's biologically plausible implementations of something similar to backpropagation (see Zipser and Rumelhart 1990 for early work) and his recent use of this type of network in motor control and so-called "mental mapping". (Rumelhart 1993b) Rumelhart's networks are primarily Hebbian but rely upon feedback from nodes producing a theoretical neuromodulator which regulates plasticity without affecting activation. (Rumelhart 1993a) The neuromodulator is plausible, but it has yet to be identified in biological systems; the present network is meant to perform a similar task without recourse to this modulator.

Rather than providing a plasticity-modifying chemical at a particular neural junction, the present strategy is simply to provide additional excitatory input to the two Hebbian nodes in question. In most networks using correlation rules to update weights, such as the example self organising network described above, high correlation between node outputs must be combined with high present output frequency to achieve a maximal update to the connection strength. Thus the strategy of providing additional excitatory input to the two nodes increases the magnitude of the connection update. Of course this will also influence the other nodes to which either of the two in question might be connected, so the strategy is far from identical to the neuromodulator scheme.

The network receives its feedback about performance from a zero plasticity subnetwork which provides a classical measure of the improvement in position caused by the most recent motor activation. A rudimentary version of such a measure is the XNOR function, implemented as shown in Figure 6 by perceptron-style units with all or nothing thresholds set to .5.

However, since we require more information than a simple verification of whether two units are either both on or both off, a more flexible measure is the function $1 - |X - Y|$, implemented as shown in Figure 7, together with its output graph. Here the nodes have zero thresholds and a continuous output response which can be read off from the cross section where the graph meets either of the two vertical planes made by the axes. I will refer to this simple network as a "convergence detector"; it may be nonlinearised by altering the output function of the final node to match the cross section of the desired graph.

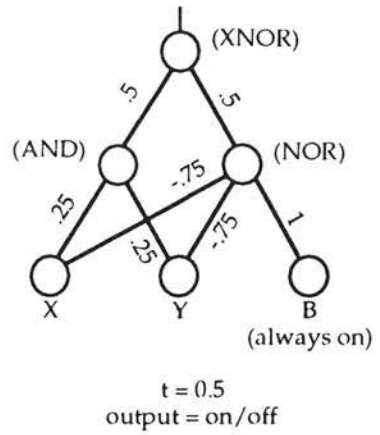


Fig. 6. XNOR function

In the complete system, depicted in Figure 8, convergence between the present image and the desired image is measured by the detector A and convergence between a previous image (thus the propagation delay) and the desired image by the detector B. While each of these detectors has a number of output signals identical to the number of dimensions of the image information, for simplicity only one output each is shown here. The outputs of the two detectors are compared by single nodes, with smooth output functions and complete efferent connections to the control network, which effectively subtract the old convergence from the new in the case of the subtractor marked C or the new convergence from the old in the case of D.

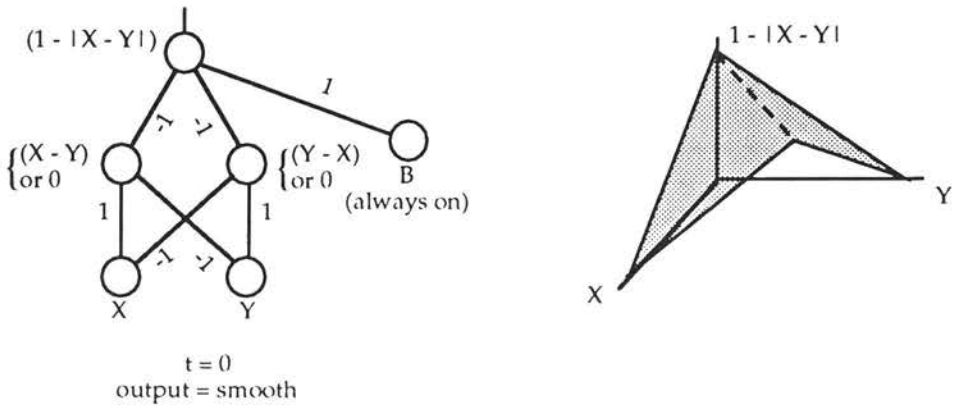


Fig. 7. Convergence detector and output graph

The output functions of these nodes must be scaled to amplify positive results: since we would expect sudden very high improvements

in convergence only rarely, it is important to magnify the presence of even minor improvements to a significant level. The D node is unique in that it represents a neuron which produces only inhibitory chemicals, but its efferent connections are still updated according to the Hebbian rule in place. (Note that some sort of arrangement with fixed afferent connections and plastic efferent connections is necessary for communication between nodes with fixed connections and nodes which are part of a learning network.)

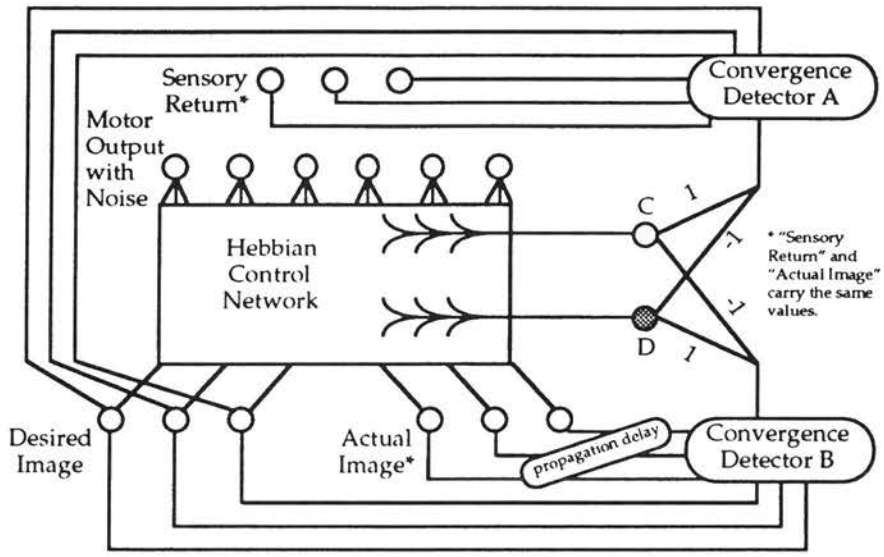


Fig. 8. Complete system

The upshot of this system is that when the controlling network yields a motor output which improves the convergence between the desired and the actual visual input, all the nodes in the control network will receive additional excitatory input from C, but only the connections with those which were firing at a high rate will be strengthened. When the motor output worsens the convergence, all the nodes receive inhibitory input from D, but again only the connections with those which were firing will be strengthened.

Additional excitatory input to a pair of firing nodes which are connected to each other in the main network will contribute to an increase in the strength of the connection between them, whereas inhibitory input will contribute to a decrease in the connection efficacy (or at least a relatively smaller increase). The interaction between the inhibitory and excitatory inputs from the subtractors themselves and the influence of this

input on the connections between the nodes within the main network is meant to be a primary mechanism affecting the system's development.

10.4.4 Every Architecture Has Its Problems

Perhaps the most telling criticism against the system as it stands is that there is no guarantee the main network will converge when it is trained upon a set of desired images with a consistent environment and consistent motor actuator characteristics. As I have stated, however, the architecture is intended primarily as an example of the kinds of features which might emerge from the design theory I described above, and it illustrates the application of a hybrid architecture to the feedback hypothesis. A more subtle criticism is that complete efferent connections from the subtractors suggest that in terms of reinforcement no discrimination will be made between nodes whose firing was highly desirable for achieving a convergence improvement and those whose firing was only marginally desirable; all nodes whose firing contributed to a positive change in convergence are reinforced, while all those whose firing contributed to a negative change are inhibited. Likewise, there is at present no mechanism for capping the influence of feedback if the network does tend toward improved performance at the motor control task; there is some danger of saturation. Improvement of the network to remedy these deficiencies and others awaits full implementation, testing, and detailed analysis.

10.4.5 Evolution and Other Goals

I have outlined what I believe is the most powerful theoretical framework for biologically plausible artificial neural network design considered to date. I have given reasons for broadening the class of neural network characteristics under the control of a genetic algorithm as well as reasons for incorporating a "learning phase" of ontogenetic development and for applying the fitness function to the phenotypes thus produced. I have indicated how this theoretical framework may allow us to view some phenotypes as hybrid networks in which evolution has "discovered" useful classical functions and embedded them in low plasticity subnetworks. I have given an example of a problem which could benefit from this type of approach and described a manually designed sample network which might be useful for solving the problem.

For complex tasks, the capabilities of networks produced under this theoretical framework may surpass those of most artificial networks either genetically created or manually designed for ordinary learning. Since artificial networks are not constrained by the boundary conditions of biological neurons operating in real space, such artificial networks may ultimately surpass even the capabilities of their biological counterparts with similar numbers of connections. Given computationally simple correlational learning algorithms and ontogenetic phases of a few thousand cycles or less, this type of neural network design is well within the bounds of existing technology.

Already the networks which emerge from simpler genetic algorithms are extremely difficult to analyse in terms of the functions of individual neurons. The architectural subtleties made possible by the framework I have described may prove still more resistant to functional analysis. This makes yet more pressing the need for ongoing consideration of the problems created by automated generation of more and more capable neural networks which we understand less and less.

10.5 Real, Artificial, and Back to Philosophical

Once again, we can see from the example network that what artificial neural network researchers allow themselves to dub "biologically plausible" is actually a far cry from biological reality (and I am scarcely less susceptible to this optimistic tendency than other writers!). But having explored the abstractions offered by these two network examples, we can now better understand some of the computational aspects of what otherwise looks like a mass of grey squishy cells connected together in strange ways with long branching fibres. In the last phase of our prelude to actual neural architectures for self models, we address briefly some philosophical considerations centring first on what we mean by 'information' when discussing information processing and data structures and second on the closely related issue of the relationship between information processing by connectionist structures such as these and by structures more obviously symbolic and equivalent to digital computers.

11

Information and How to Process It

In the preceding chapters, we have slowly been making our way through material leading up to examples of architectures for implementing some characteristics of self model data structures. Like our side trip into quantum mechanics in Chapter 4, the present chapter is a side trip to be clear on notions of information processing and to be sure that our emphasis on specifically connectionist models of the material substrate is safe from some potential criticisms from those of a classical artificial intelligence persuasion. We begin with a note on the representation of information in distributed systems before moving on to the debate between connectionists and the classical AI camp. In the final section, we conclude that whether or not connectionists generally have a case to answer, we are here safe from the criticisms of AI.

11.1 Implicit and Explicit Information

In keeping with the treatment of information suggested in the context of quantum mechanics, we shall here view information as correlations between the states of material structures. That is, if the state of one material structure is correlated with the state of another, then we say that they each bear information symmetrically about the other. When we say that information is *processed* by a neural network, we mean generally that correlations between states of activation of particular neurons or neuronal groups are created, destroyed, or otherwise transformed. Only in special cases do we mean something as narrow as the kind of information processing performed by the simple fixed plasticity classical subnetworks described above.

One consequence of this view of information is that a distributed system such as a neural network may contain information about, say, a

curve described by a particular function without containing any explicit representation either of the function or of the curve. We might be able to treat the network as a "black box" to which we feed questions such as where on the curve particular points on a line might be mapped and from which we receive answers in the form of points on the curve. Yet, if we opened up the black box and looked for some straightforward "encoding" of the function or the curve, we couldn't find it.[63] In some cases, functions may be implemented simply in neural networks with something like binary logic (as in the classical XNOR network), but in other cases there may just not be any straightforward computational implementation. In other words, a function may be stored *implicitly* in terms of the functional relationships between nodes and without *explicit* representation of any kind.

This suggests that the notion of representation to which we appealed in our initial discussions of self model properties need not be that of explicit representations, and indeed pattern matching between such representations, for instance, may take place without any sort of internal reconstruction of whatever was being represented. (See Fatmi, et al 1990 for a note on this basic point; see Clement, et al 1991 for ideas on analogue implementation.) This reveals what was misleading about the reverse Mercator projection example, for instance: in the self model, there needn't be any physical rendition whatsoever of either the external environment or the system itself, and the representations of the self model needn't be like the reverse Mercator projection at all! All we require is that relationships between the system and the environment and within the system itself be mirrored functionally in the implementing architecture. The self model view is not in any way wedded to a naïve representationalism, and our use of the term 'data structure' does not commit us to any kind of classical symbolic view of data. (For thoughts related to the superfluity of symbolic representation in pattern matching, see Gabor, et al 1960 and Fatmi and Resconi 1988. For representations and quantum ghosts, see also Marcer 1992.)

Another consequence of this view of information is that we may say a network performs *computation* without committing ourselves to

---

[63] In general we can find some encoding of almost anything in almost anything, but here we mean something systematic and straightforward which might, for instance, suffice for more than one network encoding different things. (See also Chapter 19.)

what Edelman spurns as the "instructionist paradigm", which is the view that neural structures actually compute values such as, for instance, the present angle of a joint, in order to determine an appropriate motor action or whatever. For our purposes we shall take computation as being equivalent to information processing as we have already defined it. (I am aware there is an enormous literature related to computation and computationalism and a number of incompatible definitions available, but it will suffice for our purposes to adopt this simple approach.) Thus, given the broad connotation of information which we have adopted, whenever some transformation is performed on information, we will say that a computation has occurred—even if that transformation was not "computable" on the narrow recursion theoretic definition of the term. It is absolutely essential to be clear on this use of terminology in order to see through whatever chimeric confrontations with Edelman or anyone else may seem to arise. By interpreting information as physical correlation (along the lines of Landauer 1991), we remove ourselves almost entirely from the fracas over computation, and we avoid the kinds of abuse hurled between, for instance, the Crick and Edelman camps.[64]

11.2 Classicists vs. Connectionists

Closely related to these ideas of information and representation is the ongoing debate in philosophy of cognitive science over the relationship between symbolic computation of the sort performed by digital computers and the information processing performed by connectionist networks. It will be useful for us to establish just where in this debate the correct self model view places us; in the end, we shall see that the debate is rather overblown, and our view shouldn't be too susceptible to criticisms from either camp. Battle lines for the debate were drawn most famously by Fodor and Pylyshyn's scathing 1988 attack on the

---

[64] Edelman and Crick research at the Neuroscience Institute and the Salk Institute, respectively, scarcely one mile removed from each other in southern California. But their physical proximity isn't matched by the similarity of their approaches to consciousness, with Crick advocating an ordered computational neuroscientific understanding of cognition and Edelman a disordered somatic evolution understanding. On our interpretation, a brain operating under Edelman's rules is "computing" just as much as one operating under Crick's. Indeed, even on other interpretations, there is nothing to stop an Edelman-type process occasionally giving rise to a thoroughly Crick-type structure. In general, we are concerned with what may *actually* be going on in various areas of a cognitive system rather than with how we should label this or that process.

connectionist project, and argument, parry, and riposte between various protagonists have continued ever since.

Broadly speaking, standard bearers of the classic symbolic camp—the likes of Fodor and Pylyshyn as well as McLaughlin (Fodor and McLaughlin 1990)—take positions similar to that of so-called "strong AI", whose proponents believe that all cognition is a matter of manipulation of symbol strings of one form or another. Classicists claim, among other things, that distributed systems like connectionist networks cannot achieve systematicity without actually implementing a classical (i.e., symbolic) system. Systematicity can be understood for our purposes as the capacity to generalise logical inferences. That is, once the system has learnt a pattern of inference, it ought to be able to apply that pattern to any other set of syntactically similar inputs. The challenge is often put to connectionists in terms of recognising constituent structure in distributed representations, or of recognising the logical form of what is being represented and applying transformations to it without actually extracting it from distributed form into something explicitly symbolic.

In fact, it is becoming more and more clear every month that the connectionists are winning the skirmishes in what has been called a "battle to win souls" (McLaughlin 1991) in cognitive science and artificial intelligence. Although perhaps not all the early answers to Fodor and Pylyshyn's criticisms (such as Smolensky's much discussed 1990 tensor product variable binding) were entirely successful, more recent developments, including the outstanding work of Browne and Pilkington (1994) is slowly eroding the remaining arguing points on which the classicist camp may choose to pick. (See also Chalmers 1990, Chrisman 1991, Niklasson and Sharkey 1992, Sharkey 1992.) For our purposes, however, we may dispense with a lengthy journey over the shifting ground underlying the issue and concentrate instead on what, if any, significance the debate may have for the instantiation of self models.

11.2.1 Classical and Connectionist Levels

Much of the fuel for this debate, as for so many others, comes from confusion over the different levels at which problems may be described. In truth, if we examine any classical system—in the physics sense—at a low enough level, we might interpret it as a classical system—in the artificial intelligence sense—which could in principle be simulated on a

digital computer. This is because at the very lowest level, all we are concerned with are particles interacting in accordance with the classical laws of motion. Thus, if all the classicists—in the artificial intelligence sense—were maintaining were that cognitive systems can *at some level* be understood through computable manipulations of symbol strings (with symbols representing the states of elementary particles), they would be entirely correct.[65] On the face of it, this is inconsistent with Edelman's protests that,

> "the pattern of neural circuitry...is neither established nor rearranged *instructively* in response to external influences... This is consistent with the selectionist notion that, in contrast to computers or Turing machines, there is no general-purpose animal—only the adaptive evolution of particular sensory sheets and adaptive motor ensembles and of the somatic selection principle itself evinced by particular mechanisms within the phenotype." (1989a, p. 19) [emphasis original]

But Edelman here may be understood in other than the way we've just suggested (that is, the way which holds only that classical physics can account for changes in the patterns of neural circuitry): Edelman means simply that at the level of the *patterns* of neural circuitry, the organism does not make some determination about appropriate connectivity and then undertake to instantiate that connectivity. Thus, Edelman is correct at the level of these patterns but not at the level of classical physics which supplies the actual molecular mechanisms for changing patterns. (Edelman himself is, incidentally, guilty of confusing these levels of description when he criticises Hebb for being instructionist at the cellular level.) This difference in levels reveals what is at the heart of the debate between the classicists and the connectionists, and it reveals why the debate is largely irrelevant to our project.

We can see the debate is really over the level—or perhaps, although it is rarely acknowledged, the *levels*—at which we can locate the dynamics which are necessary and sufficient for cognition. (Broadbent 1985, McClelland and Rumelhart 1985, Corbi 1993) Thus, the classicists argue, in

---

[65] This is wrong. For the moment, we'll conveniently ignore the existence of entirely deterministic but noncomputable (in the recursion theoretic sense) classical processes. (See Pour-El and Richards 1979, 1981, 1982, for instance, and also Aberth 1971.) Our analysis of levels in the debate is mostly independent of this convenient misrepresentation.

short, that the essential features of cognition are found in computable manipulation of symbol strings, even if such manipulation might actually in practice be instantiated by networks of neurons. (We must remember the powerful argument for incorporating some nod to the connectionists: when we look inside a skull, we really do find, among other things, bunches of interconnected neurons!) The connectionists, on the other hand, argue that at least some essential features of cognition can only be instantiated at the sub-symbolic level, at the level of neural network interactions which cannot be directly interpreted as representing manipulations of symbols which the organism uses. But for the purposes of locating *sensation* in a data structure, does this difference really matter?

11.3 Information, Levels, and Resolving the "Conflict"

In keeping with the observations we have just made about levels of description, we can see that the self model data structure is at a higher level of description than that where we are concerned with the type of processing going on. As long as there *is* a self model data structure, changes to it might be made by symbolic manipulation, by connectionist sub-symbolic information processing, or by little green men activating switches under the influence of some other kind of processing altogether. Thus, on the view we have been exploring, sensation *itself* needn't rely exclusively on either connectionism or classicism, because the relevant data structures might be implemented any number of ways.

Having said that, it might still be true that *human* style sensation and behaviour *does* depend on some particular mixture of types (symbolic or sub-symbolic) of neural processing because of the kinds of data transformations which they enable. Thus if, for instance, it is a property of human consciousness that sometimes transitions between conscious states are noncomputable or nondeterministic, then human systems must exploit some kind of processes which are sub-symbolic (at some level of description, of course, since they could still in general be called symbolic at the lowest level of classical physics) in order to accomplish the relevant data structure transformations. If, on the other hand, all such transitions are entirely computable, then perhaps the instantiating wetware could likewise be entirely symbolic, and human style sensation could emerge from data structures sitting atop the classical string manipulations of

strong AI. But with respect to the self model data structure, questions about essential types of processing are an empirical matter and will not be settled by rational argument from the "first principles" of what it is to be such a data structure. We will later see arguments for the *possibility* of nondeterministic transitions at the level of introspective awareness, but it is good to keep in mind that nothing of the self model view we have been developing rests either on that possibility or on any particular outcome to the classicist/connectionist debate.

With the last of these philosophical considerations settled, it is time to proceed with outlining some actual architectural examples of self model implementation. In the next section, we adopt a connectionist-inspired style to explore wiring diagrams of the building blocks of consciousness.

12

Circuits of the Self

Having explored a number of topics related to the view of the self as an abstract data structure, it is time now that we offered some ideas as to how these data structures may be instantiated by neural wetware. In what follows, we shall be concerned for the most part with exploring the kinds of architectures which may have emerged in more or less adult organisms; with few exceptions we're not offering an Edelman-style account of the cytoarchitectonic features and learning rules which enable this emergence itself. We began by postulating the self model at a sort of middle level to explain how higher level awareness might come from a lower level neural substrate, and we will now try to infer downwards in the hierarchical organisation to see how the neurons themselves could be arranged to give some of the self model properties we need. But developing a biologically plausible neuroscientific account of the inherited and epigenetic mechanisms from which these arrangements emerge is another step, and for the moment it is beyond our scope. (Having said that, we will at least endeavour to prevent our designs relying upon any features known to be inconsistent with existing empirical evidence about intercellular mechanisms.) We start with a look at very basic instantiation of the kind of compressed representation which began our exploration of self model characteristics.

12.1 Compressing and Representing

The well documented capacity of simple pattern recognition networks and self organising feature maps (SOFMs), similar to the first example of Chapter 10, to provide output keyed to specific qualities of their input data offers the most basic form of the kind of lossy compressed representation we first discussed early in our exploration of self model characteristics. The idea is that a single node or group may respond

uniquely to a particular pattern of outputs from the cells in, for instance, a sensory receptor sheet. The strengths and patterns of connection between the receptor sheet and the groups taking efferent signals from it may be such that one and only one dominant repertoire of cells responds to each relevant output pattern from the receptor sheet. This is illustrated in Figure 9, which for simplicity shows a single layer receptor sheet and two single nodes which take their inputs from the cells in the sheet. (The full extent of connections is not indicated.) As in all the examples we illustrate here, it is important to remember that the diagrams are only suggestive of the complexity of any real neural arrangement and that we are trying just to represent the most basic functionally relevant features.

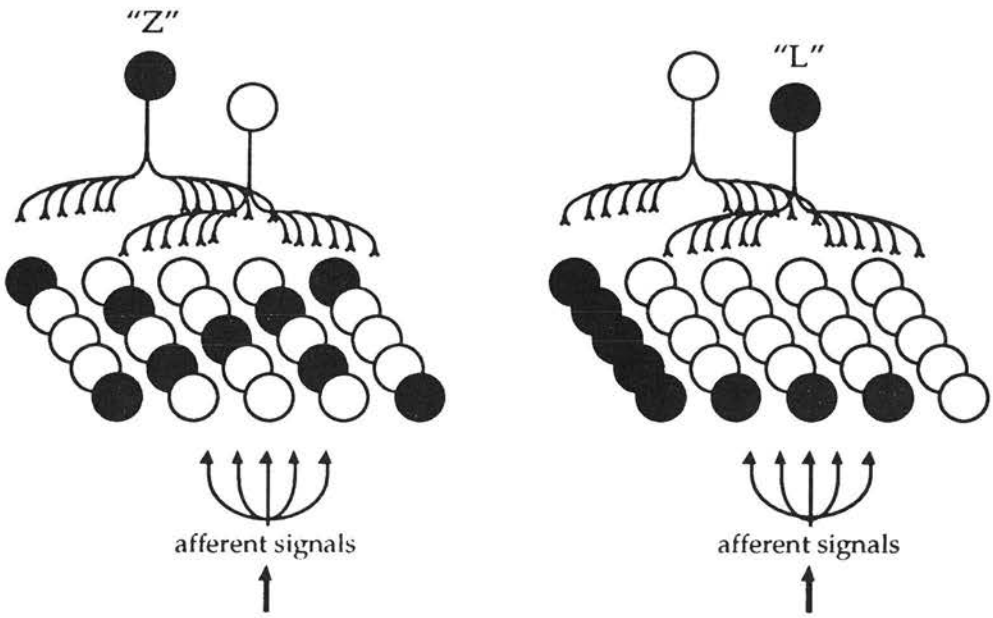


Fig. 9. Basic feature extraction and compression

In this diagram, where solid colouring indicates nodes firing substantially above base level, one upper node responds to an input pattern corresponding to an "X", while the other responds to an "L" pattern. Thus, the pattern offered on the twenty-five nodes of the receptor sheet is effectively compressed into a single node representation.

In real neural systems, there is every indication that very complex representations do not emerge in a single step, as shown here, and that instead information is combined from many different compressed representations which for our purposes can be thought of as equivalent to

very simple feature extractors. For instance, Figure 10 shows two different sheets of neurons, each of which receives the same afferent signals but which detect the presence of different features in the input data.

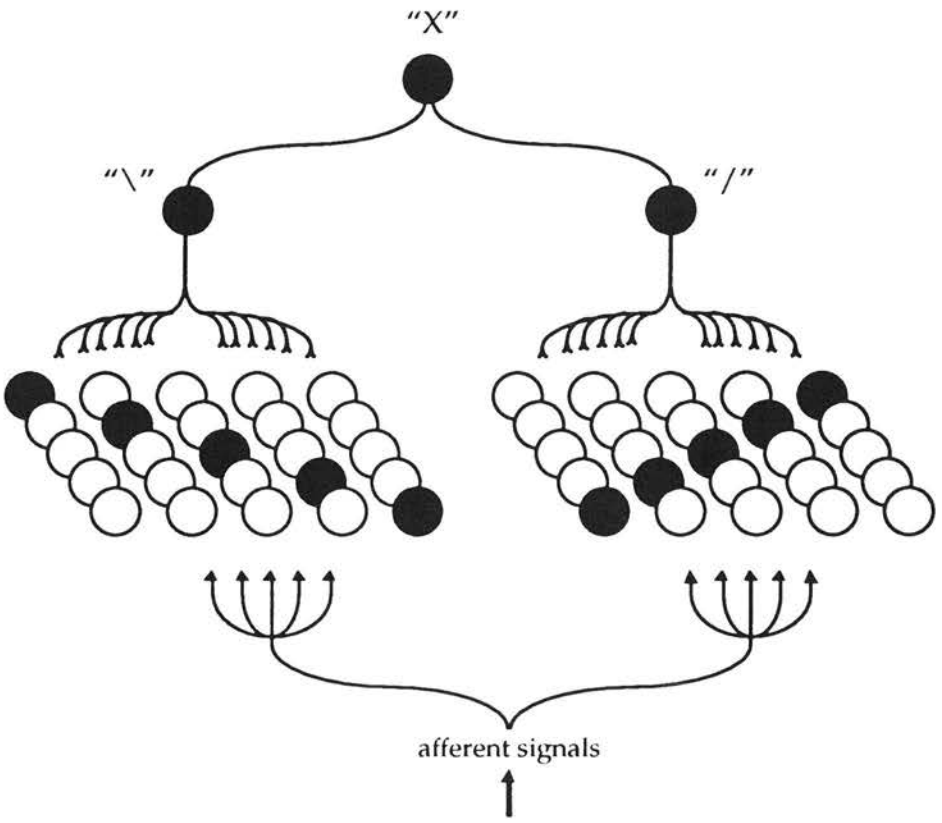


Fig. 10. More accurate compression with feature extraction

Here single nodes respond to particular patterns on each sheet and then offer their output to another node which responds uniquely to the presence of a high firing rate from *both*. This node operates much like the AND gate of digital logic. In this example, each receptor sheet might respond only to straight lines within some range of a particular orientation, so that perhaps the "L" pattern from the previous example might also be extracted and compressed by other nodes responding selectively to a vertical line from the left sheet and a horizontal line on the right together with another node which responds to a combination of these two features. (Again, note that we are here indicating with single nodes what would more likely be an entire repertoire of neuronal groups.)

Pattern recognition of more complex objects with many distinct features undoubtedly requires applying many different feature extracting

units and combining their outputs into what amount to highly compressed representations removed by several levels of organisation from the original sensory neurons in, for instance, the retina of the eye. In the course of propagation up through these levels of compression and feature extraction, signals may of course contribute to the formation of ancillary connections between groups whose outputs are frequently correlated. It is to this development of cross association that we now turn.

12.2 Mixing Company

Cross association between groups representing correlated features from one or more input modalities may be subserved through anatomical reentrance. Figure 11 shows a simple bimodal association enabled both at the level of the basic receptor sheets (where reentrant connections are shown greatly simplified) and at the level of an early feature extraction.

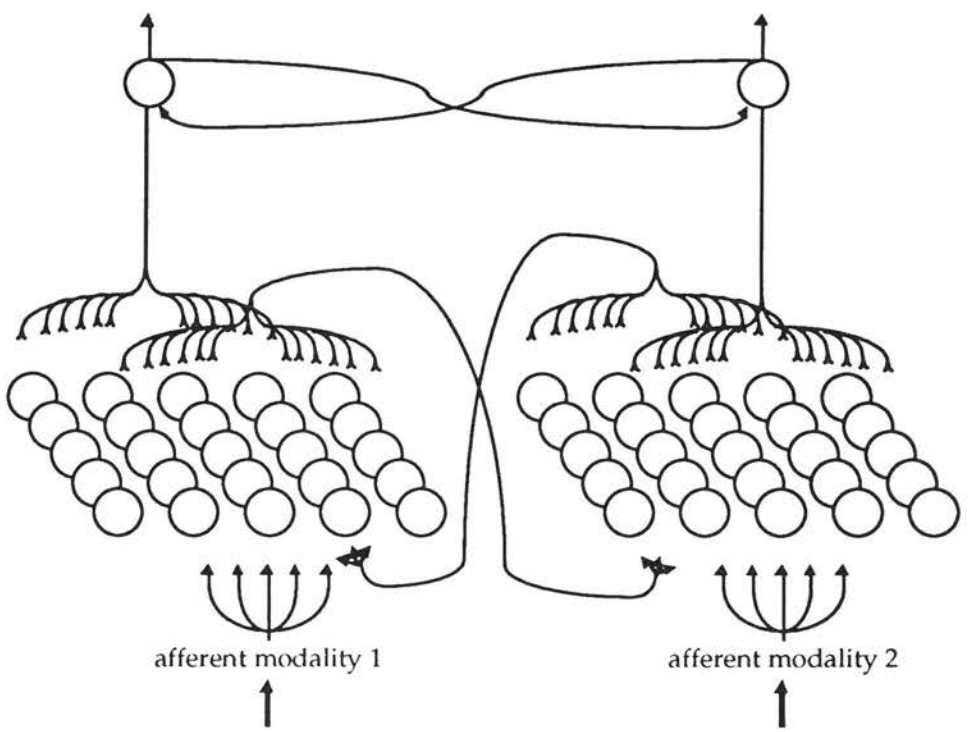


Fig. 11. Simple two level bimodal association through anatomical reentrance

Here the output signals of each receptor sheet return as afferent signals for the other, with the effect that cells whose firing is correlated in the separate sheets develop reinforcing connections with each other; thus,

presentation of a particular pattern to only one modality may raise the firing rate of some neurons on the other receptor sheet which are normally correlated with those of the first. At this level, the architecture is very similar to Edelman's own Darwin II "classification couple".

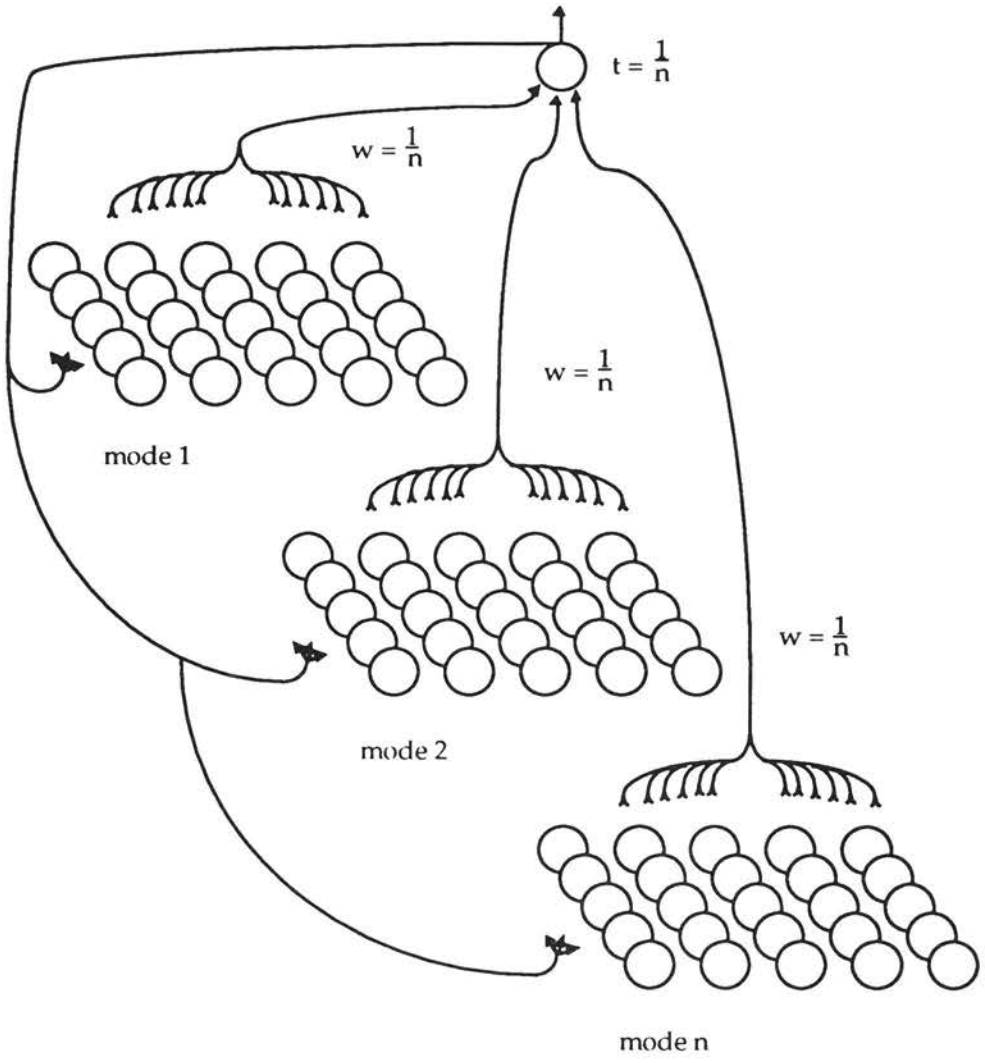


Fig. 12. Enabling low level polymodal associations without direct reentrance

This mutual reinforcement is mirrored at the higher level by the reentrant connections between the single nodes shown responding to efferent signals from the receptor sheets. The growth of these reinforcing connections may be promoted by mere physical proximity and a variant of Hebb's rule (or other correlation rule) or in some cases even by the development of gap junctions or ephaptic interactions.

A simple variation on this theme emerges when we dispense with the intermediate feature extracting nodes and provide a single node at a higher level which responds to features from any of several modalities and then propagates feedback to the much earlier sensory sheets, as in Figure 12. Here the synaptic weights of the inputs to the high level node are matched with its own threshold to ensure that the presence of any single one of the relevant input patterns is sufficient to activate the high level representation and the feedback signal.

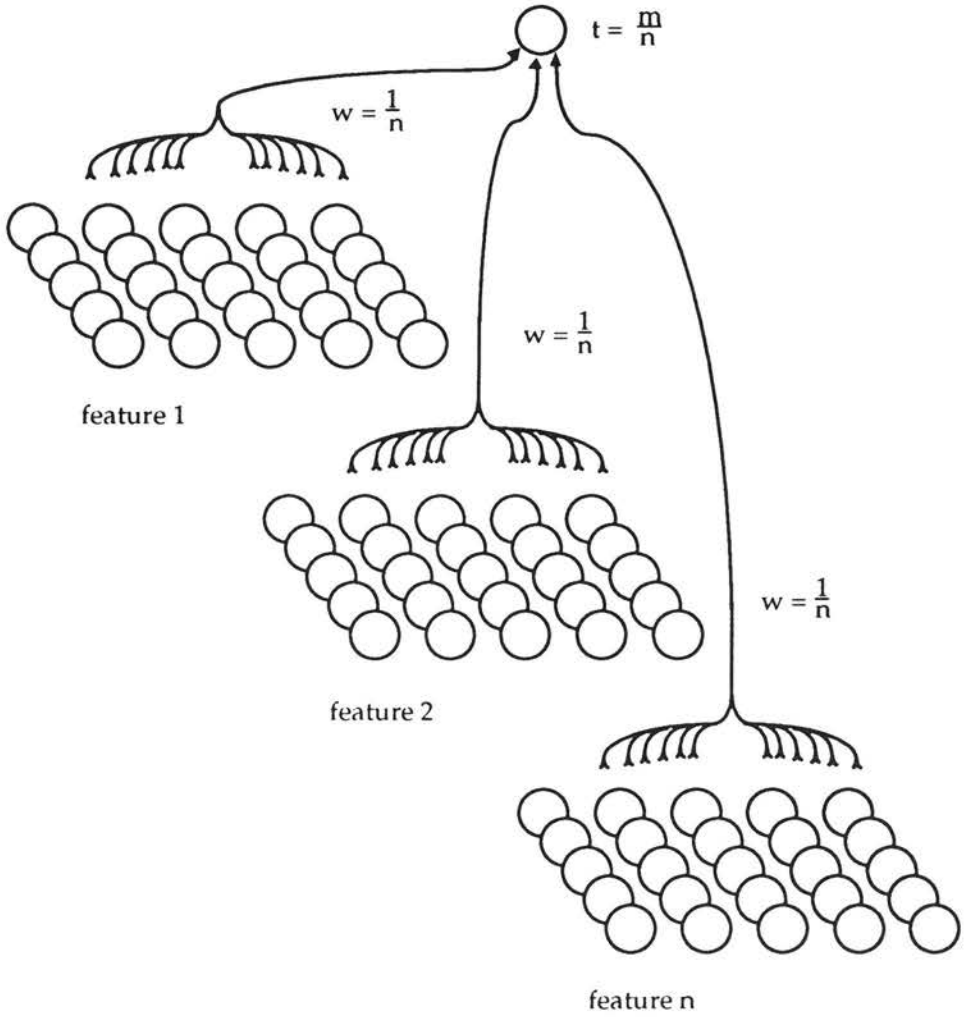


Fig. 13. Polymorphous categorisation for m out of $n \geq m$ properties

This arrangement has the advantage that even without the kind of physical proximity between the receptor sheets which would allow them to develop their own reentrant connections, a higher level feature extractor and compressor could still promote cross association at the lower

levels with simple feedback. This is one application of the rôle of self model components in overcoming the physical limitations on direct neural connectivity.

A similar network, without the presence of feedback and with a slightly different relationship between weights and threshold on the highest level node, can, incidentally, instantiate a simple kind of polymorphous categorisation like we discussed in the context of neural Darwinism and perceptual categorisation. Figure 13 shows an arrangement for detecting the presence of any m out of $n \geq m$ possible characteristics from receptor sheets each responsive to different types of these characteristics. (For simplicity, the likely intermediate nodes taking input from each receptor sheet and providing their outputs to the highest level node are not shown.) The relationship between the weights and the threshold shown guarantees that when m or more relevant features are detected, the highest level node will fire.

Combining the sort of arrangement of the last few examples with the feature extraction of Figure 10 and the higher level structural reentrance of Figure 11, we see in Figure 14 an example of higher level compressed representation from a set of three mutually reinforcing feature extractors which can together be thought of as a sort of "correlation extractor", strikingly similar to Hebb's unimodal cell assemblies. For the simplest cases, it is something like this type of architecture which we should expect to underlie the polymodal representation which we encountered earlier with respect to the development of symbolic language.

Here, three nodes, each responding to receptor sheets taking afferent signals from different modalities, have developed mutually reinforcing connections as a result of correlations in their firing frequencies when the same kind of source stimulates all three modalities simultaneously. As we saw in Figure 11, these connections can encourage firing in neurons primarily responsive to output from single modality receptor sheets even in the absence of appropriate signals from those sheets. In the network depicted in the present diagram, when one or more of the three initial single modality feature extractor nodes is excited to a sufficient level that the other two nodes with which it is correlated are also enticed to fire—and notice that once this begins to happen, the signals from the newly excited nodes feed back into the system of three so that the firing tends to maintain itself—the synaptic weights and the threshold of the final node

are matched such that it, too, will fire. But in the absence of sufficient signals from all three, the firing frequency of the highest level node remains at its base rate.

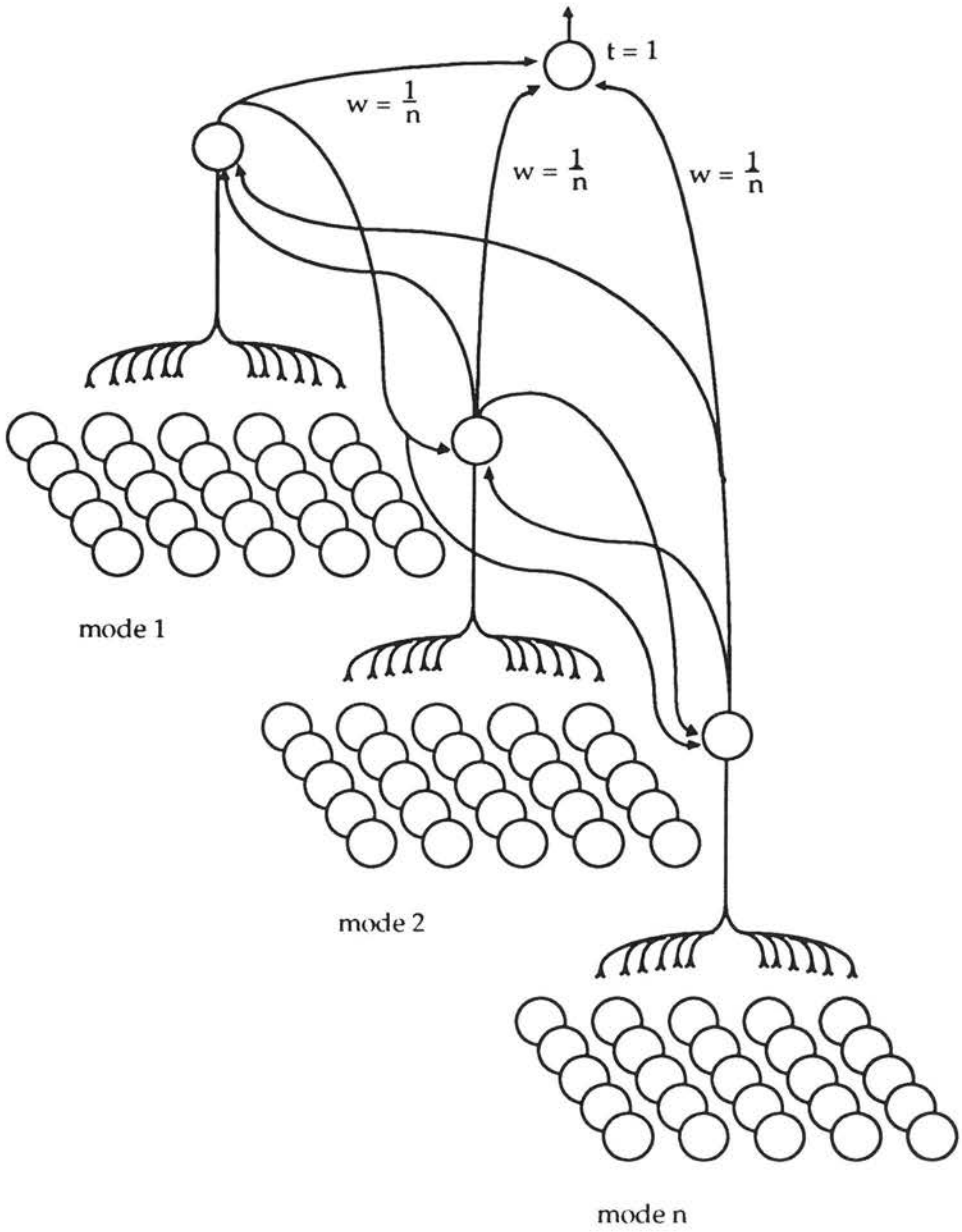


Fig. 14. Cross association for polymodal representation

A real implementation of such a network would of course involve multiple feature extractors, each sensitive to particular kinds of output from each of these receptor sheets. As a result, the reentrant connections between the populations of neurons here represented as the single set of

actively researched by cognitive neuroscientists, and we will here discuss a couple of the possible remedies.

12.2.1 Freeing Our Inhibitions

One common suggestion is to incorporate "interneurons", or nodes producing inhibitory chemicals which sit between an excitatory neuron and some of the cells to which it would send inhibitory signals. Thus, subsequent nodes could receive EPSPs, or excitatory post-synaptic potentials, from the main neuron and IPSPs, or inhibitory post-synaptic potentials, from the interneuron. Two such neurons, one providing inhibition and the other excitation, might even sit just next to each other, such that gap junctions or ephaptic interactions synchronised their output frequencies. Alternatively, receptor sheets and other structures in the architectures we've seen may just be peppered with a mix of inhibitory and excitatory cells, and connections might simply develop in a way roughly functionally equivalent to the architectures we've shown.

My own suspicion is that an excited cell releases chemicals which tend to inhibit other physically proximal cells that are not directly connected to it. Individual cells might even be "tuned" for differential responses to particular neuromodulators so that, for instance, nodes in a particular population were excited by the firing of one neuron, while nodes in another were inhibited by the same neuron's firing. However, such a neuromodulator remains at this point purely theoretical, and the search for mechanisms to subserve the functional equivalent of mixed IPSPs and EPSPs from the same neuron continues actively.

Returning to the kinds of architectures which prompted our side comment on inhibition, it is of course true that compression and representation needn't always be of sensory receptor sheet output, as we have shown here. There is nothing to prevent similar architectures from developing around compression and representation of relationships between, say, sensory receptors and motor actuators, or between any two or more separate neural assemblies. Whatever is being represented, the kinds of higher level cross correlation we've described enable groups in the higher levels of the organisation to mirror the functional relationships between groups at the lower level, just as we originally indicated the self model requires. But to act as other than a passive

mirror, the capabilities of these higher level representations must be exploited in such a way that the information they contain can be usefully processed and fed back to lower level assemblies. It is to this that we turn our attention in the next section.

12.3 Control Systems—Self Models on Top Again

We begin with a simple "digital" example of a shared bus architecture for three separate classical processors. The idea of a shared bus common to several processors is common in parallel computing. However, the shared bus is normally used only for such a purpose as feeding each processor the same stream of instructions; typically a separate bus gives each processor access to its own portion of memory. A different rendition of a shared bus is pictured in Figure 17.

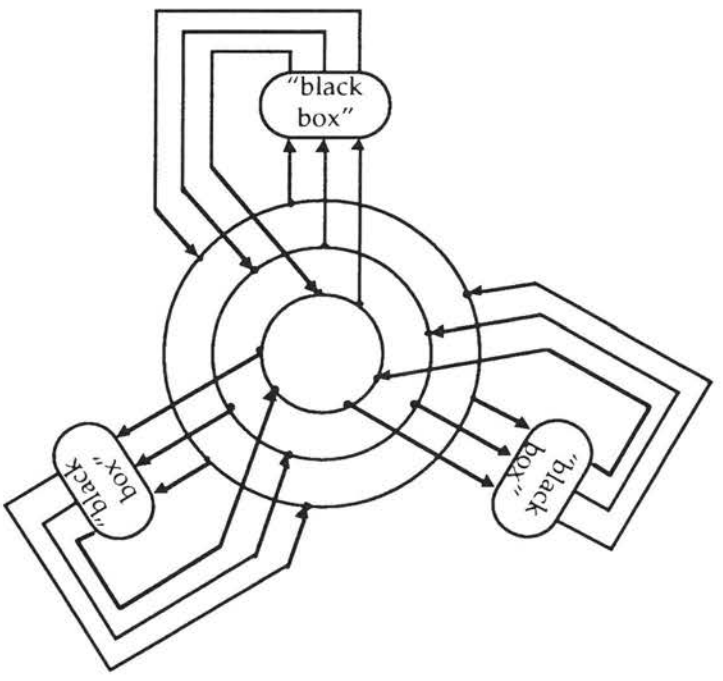


Fig. 17. "Digital" shared bus ring architecture

Here, each processor ("black box" because we needn't know what goes on in any of them) is offered only three inputs, and each processor's three outputs is fed right back to the same three lines which provided the inputs. At least two different ways of handling the input lines could be at work in each of the three black boxes. On the one hand, they might just

look to some subset of the lines as an "activator", such that a processor was activated only if it saw, for instance, a 1 and a 0 on two lines. The other line might be used for the processor to provide its output. Since another processor on the shared ring bus might actually look at different lines to determine its own activation, the output of one processor could possibly be on to an "activator" line of one of the others. Alternatively, all the lines might be be used by all the processors both for sending their outputs and for determining their own activation.

A few minutes spent contriving different sets of logic gates to replace each of the three processors (in addition to simple refinements like clamping lines low in the absence of a high signal) reveals how easy it is to create rings whose processors are activated in any desired sequence, and with the addition of input-dependent variable propagation delays before output signals become available, it is possible to create very complex (probably even chaotic) inter-processor dynamics.

There is some reason to believe that such an architecture for exploiting a non-multiplexed bus shared between independent processors could be highly useful for some parallel computing applications, and as far as I am aware no such architecture has yet been reported in the relevant literature. In general, however, the typically synchronous nature of today's parallel computers is not well-suited to such a shared bus—although the addition of proper wait states and dedicated control lines indicating when a processor was "ready" to provide output and so forth might remedy some of the most obvious problems.

What might be better suited to this type of shared bus, however, is the kind of massively parallel, reentrant, asynchronous architecture common to neural systems. In neural systems, there are no address or data buses, and there are no timing signals or wait states ensuring that outputs will only be provided when neighbouring neurons are ready for them. In general, the requirements of neural processing are wholly unlike the requirements of digital processing, and where digital computers are not immediately suited to such a ring architecture, we have actually already seen something like this architecture at work in our earlier diagrams, although we did not note it at the time. Figure 18 depicts a neural ring architecture, and while the connection strengths might be different, a quick mental warping and tearing reveals the structure is connected like the correlation extractor we saw previously in Figure 14.

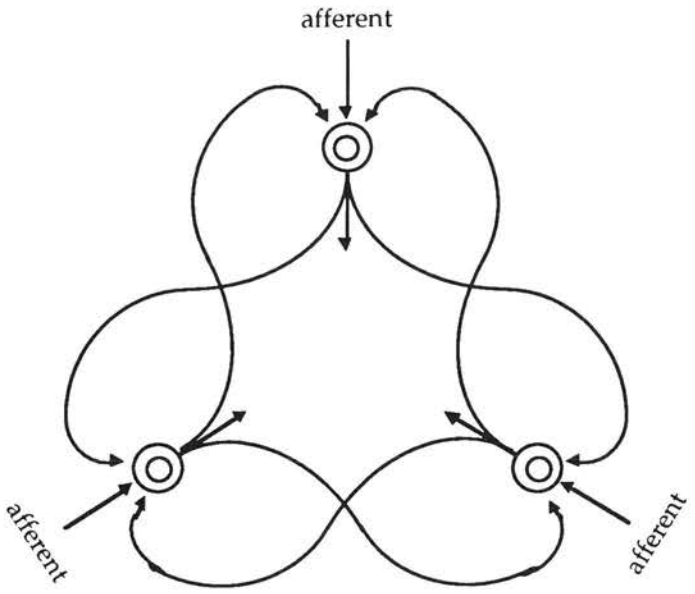


Fig. 18. Neural version of the ring architecture

Here, single nodes are replaced by a double ringed "node" which indicates simply that we are almost certainly concerned now with large neuronal repertoires rather than single nodes or even smaller groups. Although we made this *caveat* earlier that wherever we illustrated single nodes we should think of groups, here it is even more likely that such control structures would be handled by large reentrant populations.

It is this type of simple control structure which may be one of the central and most important architectures underlying the complex information transformations in the self model data structure. It is from these types of structures that we might expect output reflecting a "simulation" of a set of conditions, and it is from these types of structures that we might expect the kind of downward influence which we suggested should be possible from the self model back to lower level assemblies.

The motivation for the first claim is the observation that with appropriate relationships between synaptic efficacies, the behaviour of the kind of "correlation extractor" we explored before becomes useful as a simulator. That is, if there is information in the system (in the form of output frequency relationships between nodes in the correlation extractor or in the ring) about, for instance, what sensory input would be correlated with a particular motor output, then a neural group *representing* the

activation of that motor output could simply be stimulated by another signal and the influence on the sensory input read off from the effect of that representation's activation on the group representing the activation of the particular correlated sensory input.

The motivation for the second claim derives directly from this: if simulations such as this *can* be carried out at a high level, and there are good indications that they could be, then there is every reason to believe that the "results" of such observations could be usefully offered back to the lower level assemblies. One way in which such output might be used is in the kind of "supervised" reinforcement learning we saw in the chapter on artificial neural networks, and another might be in "opening a channel" for two physically separated groups of neurons to communicate. We might picture the latter as something like a neural interchange hub—related to the abstraction window of earlier chapters—as shown in Figure 19, where the disinhibitor nodes take afferent signals from some other group, perhaps from some one of the groups in a neural ring architecture.

In this diagram, the efferent signals from input processors are prevented from reaching the interchange area by inhibition from a node which is kept firing at a sustained rate by input from a "bias" node, or a node which is itself firing at a high frequency. (Yet again, both of these rôles would likely be filled by larger repertoires of neurons.) Likewise, output from a set of nodes taking afferent signals from the interchange area is kept from reaching the output activators by other inhibitor neurons. Exchange of information between the two sides is enabled by disinhibitors, which are nodes that provide strong inhibitory signals to the nodes which are themselves normally inhibiting the exchange. When the firing rates of these latter are lowered, both the input "providers" and the output "customers" are effectively connected. Any such interchange in a real system would be an area of very high axonal and dendritic arborisation; there is some reason to believe the hippocampus might be one candidate for this type of "neural switching station".

One way of combining this sort of neural interchange hub with the simulation capacity of the ring architecture is indicated in Figure 20, where efferent signals from a sensory processor are routed from one ring assembly to another, where a "simulation" takes place before a number of signals are combined to allow a given output to take place or to open a given channel in an interchange hub.

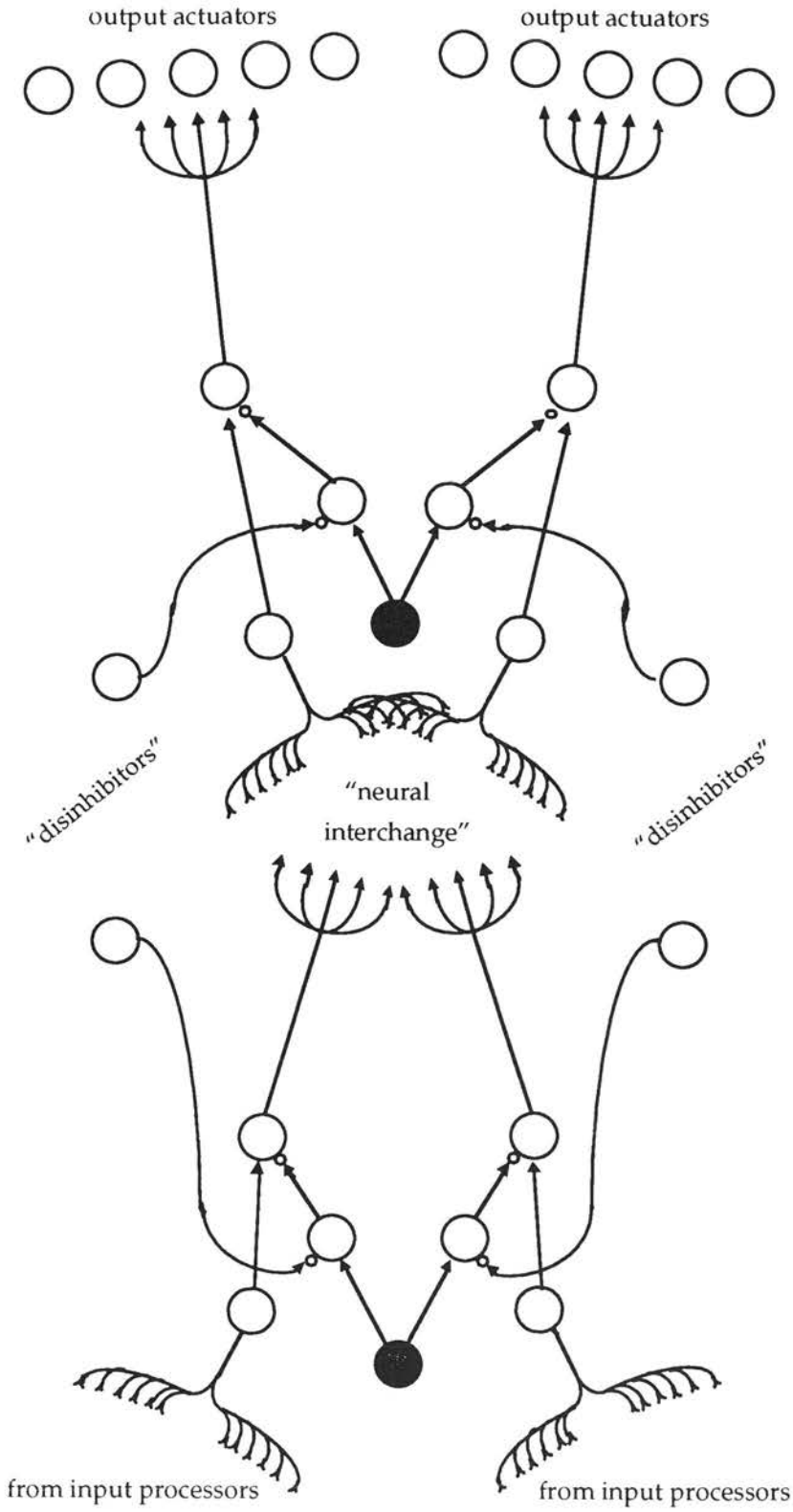


Fig. 19. Neural interchange hub—one gateway to the abstraction window

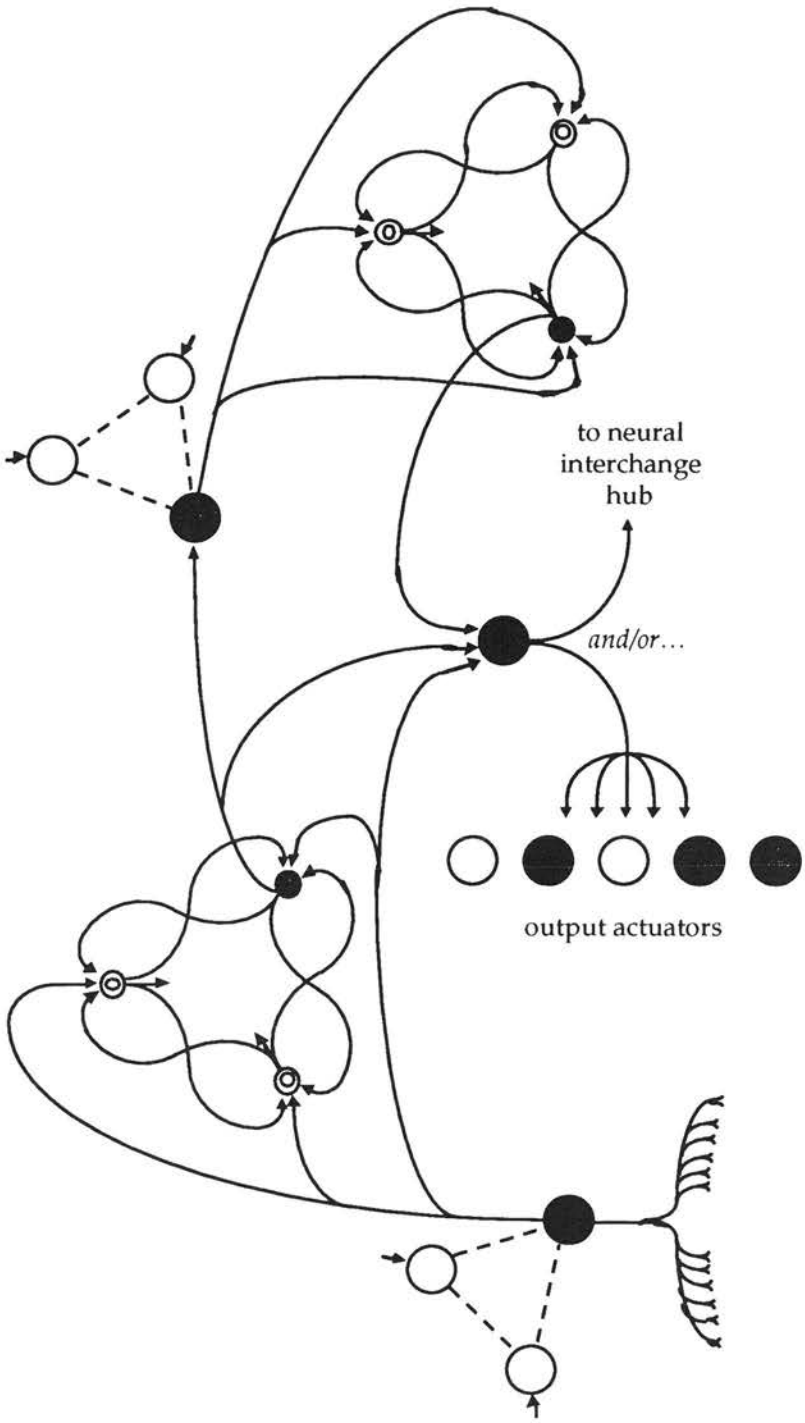


Fig. 20. Simple simulation and decision circuit

In this diagram, activation begins in the lower right corner, where input signals excite some compressed representation which is perhaps a member of a simple mutually reinforcing group or correlation extractor of the type we saw back in Figure 11 or Figure 14 (here indicated simply with

dashed lines). Efferent signals from the representation propagate into one ring structure, where they excite a particular group that provides signals to another member of a basic mutually reinforcing group, whose efferent signals in turn go to another ring in an arrangement mirroring the start of the process. (Although it is not depicted here, we can imagine the propagation of signals from the first ring to the second member of the mutually reinforcing group being mediated by a neural interchange hub and enabled by some still higher level processing or perhaps some parallel result of the activation of the initial compressed representation.) One of the groups in the second ring is activated by this input, effectively offering a "simulated" response to a "hypothetical" input as if the output from the second mutually reinforcing group had come from the environment directly. This output is finally combined with that of the first ring structure and that of the original input processor by a group acting as an AND gate, such that if there really is appropriate input *and* a particular signal from the first ring structure *and* a particular signal from the simulation, then the relevant output will be activated or the relevant channel opened in an interchange hub.

12.4 Self Circuits—All Together Now?

Of course, this particular pattern of interconnection and sequence of activation is entirely contrived and suited primarily for illustrative purposes, but the general idea of these kinds of information processing by simple interconnected neural groups is straightforward and highly capable. We have seen in this last network the possibility for sophisticated interaction between subsystems serving different input or output modalities and for the rudimentary testing of responses to stimuli, mediated by the competing outputs of structurally reentrant neural networks. Previous example networks illustrated other applications of reentry, recursivity, and heterarchical organisation in building functional representations of inputs and the correlations between them and between them and outputs.

These networks illustrate in simple ways some of the methods which Nature might have used to instantiate in neural wetware the kinds of capacities we have been attributing to self models. I am not aware of any glaring biological implausibilities among the features on which the

networks rely, and on the contrary they have much to recommend them as examples of limited plausibility. They are compatible with existing neuroscientific data and with Edelman's theoretical frameworks, and they help to bridge the gap between his low level accounts and our own exploration of higher level features of subjective awareness. On the downside, these example networks are almost certainly too uniform and simple to be found in real neural systems, and undoubtedly Nature will have found cleverer ways of accomplishing many of the things which these examples might appear to be the simplest way of doing.

Overall, however, our task in this chapter has been merely to offer some ways in which the characteristic capabilities of self models which we have explored so far might be instantiated by the brain's wetware, and in that we have been successful. With more detailed work on these types of networks, I believe it will be possible also to make more headway against the so-called grain problem and frame problem.

Our earlier limited response to the first problem was to note that the data structure is "blind" to its instantiating material substrate, and these example networks show more clearly how it is that higher level representations lose information about that which they are representing. More careful analysis of these networks will, I believe, show that no self model instantiated by discrete components can be "aware" of the granularity of any more than some small portion of the overall system and that this awareness cannot *itself* be grainy. That is, while it might be possible to contrive a circuit to allow the system to become aware of a perceptual limitation caused by the grain of its discrete instantiation—although there is little evolutionary reason to believe such circuitry would have emerged in real organisms—I believe that the awareness itself of this fact will necessarily be impervious to grain.

On the second problem, there is good reason to believe that with the development of more sophisticated combinations of the kinds of networks described here, the frame problem will actually, rather fortuitously, disappear. I believe the frame problem is a mere relic of the symbolic processing roots of artificial intelligence and that it results from trying to impose an existing logical structure—typically that of the propositional calculus—on a body of data and then trying to process it in a way resembling the processing of real organisms (the top-down approach), rather than allowing the data itself to be fitted into a system of

representation and interaction essentially of its own neural creation as is actually done in real biological organisms (the bottom-up approach). When more complex networks are developed, perhaps with the combination phylogenetic-ontogenetic scheme described earlier, I believe the frame problem will not appear for the new systems created. (After all, no network so far developed has ever faced the frame problem itself; we have only created the frame problem for our own understanding of contrived systems when we have tried to develop them into tools for doing what we want.)

So, even the questions we have not directly addressed at length perhaps have at least a bit more direction now. Having explored in this chapter and previous ones the self model approach at both higher and lower levels and discussed some of the philosophical issues on which they bear, it is time now to move on to a range of questions relevant to the dynamics of the self model's instantiating hardware or wetware. While we have so far concentrated on the relationship between the first person and third person perspective and how the apparent gap between them may be embraced and better understood, in the second half of this dissertation we will concentrate more on observations about neural systems from the third person perspective, inferring upwards to what effects characteristics of these neural systems might have on experience. Chaos will play a central rôle as we explore dynamical properties of neural networks and see how they might bear on the time evolution of the mental properties they are instantiating. We will have something to say on complexity and the relationship between chaos and noise, and in the end we will return to the issue of representation and how chaotic dynamics may bear on it.

13

The Spaces Programme[66]

In the early chapters of this dissertation, we explored the idea that the seat of conscious sensation—the seat of the self—was an instantiated data structure called a self model, implemented by some kind of hardware or wetware system, the representing and controlling of which was a central task of the self model itself. We noted some of the selective advantages with which an organism equipped with a self model might be endowed, and through some simple artificial neural network architectures we saw ways in which real self models might be expressed in distributed systems. We now turn our attention to the relationships between the dynamic properties of the self model data structures and their instantiating material systems to see what influences characteristics of the low level dynamics may have on higher level selves. In a digital computer, for instance, the very lowest level dynamics—at the level of quantum descriptions of semiconductors—are probabilistic, while whatever data structures may be implemented at higher levels are entirely deterministic and describable by computable functions.[67] Self model data structures, however, are at least less straightforwardly deterministic products of their underlying wetware than their digital counterparts.

This may be especially true when it comes to neural substrates whose dynamics are specifically chaotic. In much of what follows, we shall be concerned with the importance of specifically chaotic dynamics for intelligent systems implemented by neural networks. While its specific rôle is controversial, the capacity of some biological and artificial neural

---

[66] Chapters 13 to 19, inclusive, have been available in different draft forms on the International Philosophical Preprint Exchange and mirror sites for several months. I am grateful to readers across the world for insightful comments on those early drafts.

[67] Some data structures might not actually change deterministically with respect to information available at the same level of description as the data structures themselves, but they would nonetheless be entirely deterministic and computable with respect to information available at the level of individual logic gates.

networks to exhibit chaotic behaviour is well established.[68] Perhaps crucially, the sensitive dependence on initial conditions of chaotic systems suggests that minute perturbations in the low level dynamics of neural networks could possibly evolve over time into influences on those systems at grosser levels of description. In particular, it might be that microfeatures of low level processes which are not available to the coarse grained introspective awareness of the self model could over time evolve into significant influences on higher level features of brain dynamics which may be available to introspection.

We will explore a new representational schema which provides an economical means of formulating the interactions between dynamics at low, intermediate, and high levels of description. This scheme is similar to the influential framework of Marr (1982), but as Horgan and Tienson (1993) suggest, Marr's treatment is not immediately hospitable to the kind of connectionist context with which we are here mainly concerned. Like their framework (Horgan and Tienson 1993, Horgan and Tienson in press), our representational schema is better suited than Marr's to exploring the kinds of questions we will here set ourselves. After a very brief introduction to the terminology of dynamical systems and a short but somewhat technical look at chaos theory, we outline the representational framework and consider some of the insights we might gain from it.[69]

13.1 Dynamical Systems

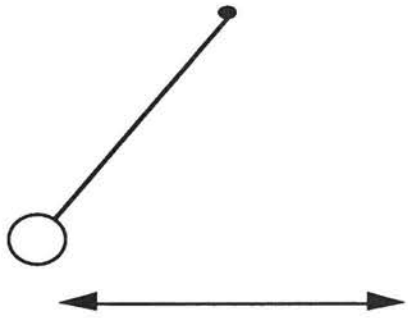
The phrase *dynamical system* has earned a place for itself in the literature of cognitive science and the philosophy of mind. Any classical mechanical system can be represented in the *state space* framework of dynamical systems theory. The state of a system can be described as an ordered n-tuplet, which fixes the value of each of the system's degrees of

---

[68] Research in this area is too extensive to list within the paragraph; to name a few of the highlights: Choi and Huberman 1983; Sompolinsky and Crisanti 1988; Li and Hopfield 1990; Kolen and Pollack 1990; Ambros-Ingerson, et al 1990; King 1991; Wilson and Bower 1992; Pollack 1992; Chapeaublondeau 1993; Fan and Holden 1993; plus the olfaction work of Walter Freeman and colleagues (cited in Chapter 7 *in toto*).

[69] Some material in this section is based very loosely on an earlier paper (Mulhauser 1993c) in which the representational schema I will describe was set within a context of fuzzy mathematics. In the present approach, I have adopted a cleaner mathematical framework in which fuzzy logic is rendered superfluous by observations about the topological relationships between the representational spaces.

freedom.[70] The evolution of a system can be represented graphically as a *phase trajectory* through n-dimensional phase space, a curve showing how each of the n variables changes with time. Aside from being a convenient way of representing classical physical systems, phase trajectories allow us to make geometrical observations which might be missed were the system's evolution represented simply as, say, columns of numbers. For instance, it is much more difficult to get an intuitive feel about the dynamics of the pendulum represented in Figure 21 from the table of numbers than from the phase trajectory depicted in Figure 22. In the second figure, we can see that the pendulum is fairly uniform on each swing except that it is gradually winding down under the influence of some damping force.



time	displacement	velocity
0	-5.0	0
.5	-4.6	.3
.9	-3.1	.8
1.2	-.2	1.3
	...etc.	

Fig. 21 Time evolution of a pendulum

The phase space framework needn't be restricted to representations of real physical systems. We might also represent the evolution of something like public confidence in a government relative to a national inflation rate. Here we might find either a continuous dynamic evolution—when inflation goes down, confidence goes up by some related amount and vice versa—or we might find discontinuous dynamics such as a sudden jump in public confidence as soon as the inflation rate reaches a particularly low threshold value. In some cases there is a clear sense in which the dynamics of a system represented at one

[70] The relationship between the number of degrees of freedom and the number of variables n required to fix the state of the system in each of those degrees of freedom varies with the dimensionality of the example and according to whether the system is conservative or dissipative.

level are indeterministic despite the fact that those dynamics are based upon deterministic dynamics at a lower level.

For instance, consider a three dimensional phase trajectory[71] indicating the relationship between the number of marine research vessels a country has at sea, the number of total marine scientists at sea, and the total funds being devoted to the scientists-at-sea programme. We might notice that in general, when the funding goes down, so does the number of vessels and so does the number of scientists, and vice versa for an increase in funding.

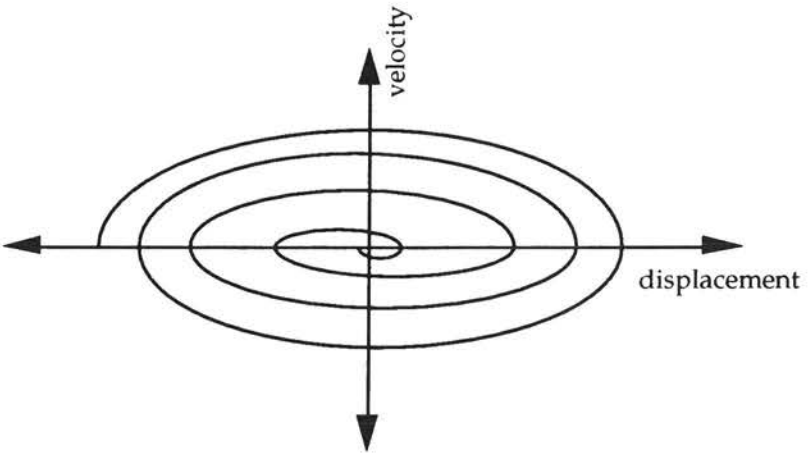


Fig. 22 Evolution of a pendulum in phase space

But the number of marine research vessels is a *coarse-grained* look at a country's marine research program; a country might have five vessels, each of which carries 100 scientist, or it might have fifty vessels, each of which carries only three scientists. These vessels might also demand different degrees of funding to stay operational. Thus, a country might have a mix of vessels such that sometimes the phase trajectory shows a transition from 94 vessels and 451 scientists to 89 vessels and 430 scientists (with a particular decrease in funding); but on another occasion there might be a transition from 94 vessels and 451 scientists to 89 vessels and 475 scientists (with an *identical* decrease in funding). If we had lower level information about the vessels themselves and the number of scientists they carried and the cost of keeping each at sea for particular lengths of

71 Since the variables here are discrete, this isn't the best example for plotting an actual phase trajectory, but the point we are making isn't compromised by this fact.

time, then under the reasonable assumption that there were some mathematical relationship between funding level and the number of scientists a country wanted to have at sea (and assuming there were no other confounding variables), we could predict the impact of any particular change in funding level on which boats would be kept at sea.

But because we lack this lower level information, the dynamics at the higher level are not predictable, even for given changes in funding. There is nothing mysterious about this conclusion: the dynamics at the level of boat numbers and scientist numbers and funding might not be deterministic on the basis of information available at that level, yet the dynamics at the level of specific boats with specific costs and specific scientist-carrying capacities could be entirely deterministic and predictable. Ignoring for the moment the significance of the λ labels, we can see in Figure 23 that even if the two trajectories depicted are entirely deterministic at a low level, their evolution is nondeterministic at the level of description of grid squares: the two trajectories begin in the *same* grid square but evolve (deterministically, at an appropriate level) into two *different* squares. It will become progressively clearer in our discussion that this emergence of indeterminism with respect to a particular level of description is typically symptomatic of that level's ignoring relevant properties at a lower level where dynamics are deterministic.

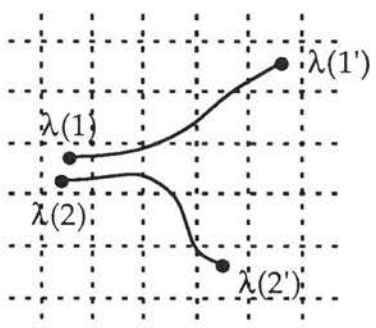


Fig. 23 Determinism depends on levels of description

Shortly we shall explore the application of the dynamical systems framework to states of mind for intelligent systems, but first we take a brief sojourn into chaotic dynamics, a particular kind of behaviour which some

dynamical systems display that can make prediction at all but the very finest grained levels highly problematic.[72]

13.2 Chaos, Graining, and Prediction

The formal definition of a chaotic system is still open to debate, but for the present purposes I appeal to one widely used definition which requires that three properties hold: the set of periodic points in phase space is dense, the system is topologically transitive, and time evolution is sensitively dependent on initial conditions. (I adopt the terms of Devaney 1988; Bergé, et al 1984 and Barnsley 1988 are similar.) The first property is the most straightforward: the set of periodic points in the phase space of a chaotic system at any given time slice (or alternatively, in any Poincaré section) is at least countably infinite, and between any two points in the set we can always find another. Periodic points are just points which lie on a closed trajectory; that is, if we ignore the time dimension, the system will keep visiting the same physical locations in phase space. Some systems include dense coverings of specifically repellent periodic points. A repellent point is simply one from which all points in a local neighbourhood diverge over time. Alternatively, if the system is invertible, trajectories in a local neighbourhood approach a repellent point asymptotically under reverse time evolution. Many strange attractors are densely covered with such repellent periodic points, and in some systems, such as the hyperbolic toral automorphisms, the entire space is so covered. Later we shall be concerned particularly with that subset of chaotic systems which include—at least in some neighbourhoods—a dense covering of repellent fixed points of all possible periods.[73]

The second characteristic which we shall take to be a necessary condition of chaos—topological transitivity—indicates that any particular neighbourhood in phase space will eventually be visited by the phase trajectory of some point lying within any other arbitrarily small neighbourhood. Topological transitivity is the property that for any f: J $\rightarrow$

---

[72] The material about chaos is reprinted or adapted from a small section of Mulhauser (1993d—please see the Appendix for reprint information); this paper receives more attention in Chapter 15.

[73] A fixed point of a given period t is the same as a periodic point whose period is t. The term is used because if the system is sampled every t units of time, then a periodic point with period t is fixed, or invariant.

J, where J is a metric space, and for any two open bounded sets U, V $\in$ J, there exists an integer n $\geq$ 0 such that

$$f^n(U) \cap V \neq \varnothing$$

In other words, the phase trajectory under f of at least one point in U will intersect V in a finite amount of time, *regardless* of how small or how far apart U and V are to start with. To put it still another way, given any two open bounded sets in the space of a chaotic system, there will always exist some finite phase trajectory connecting some point in one to some point in the other.

The final property of chaotic systems, sensitive dependence on initial conditions, or SIC, can actually be derived from the first two (Banks, et al 1992)[74], but it is sufficiently curious to merit its own explanation. SIC appears when for any state in the phase space of a chaotic system, there is another state within an arbitrarily small neighbourhood which lies on a phase trajectory diverging from that of the first. I give here the definition for the discrete case. A function f: J $\rightarrow$ J, where J is a metric space, is SIC when there exists a $\Delta > 0$ such that $\forall$ x $\in$ J and for any closed neighbourhood N of x, there exists a y $\in$ N and an integer n $\geq$ 0 such that

$$|f^n(x) - f^n(y)| > \Delta,$$

where n represents the number of times the function is iterated. (The case for continuous systems is directly analogous, with a term similar to n representing the time parameter.) Note that this property occurs for every x $\in$ J. Moreover, every neighbourhood N of x includes at least countably infinitely many diverging y. For one constructive proof, consider any such neighbourhood for an arbitrary x. The definition guarantees us a y $\in$ N whose phase trajectory diverges from that of x, so note this y and consider a new N' which is the previous N minus the y (or, alternatively, a new N' which is the previous N minus a neighbourhood around y in the limit as the radius of that neighbourhood r $\rightarrow$ 0). We are now guaranteed a diverging y' in this new N', which in turn generates a new N'' and a new y'', ad infinitum.

In typical chaotic systems—if there is such a thing as a typical chaotic system—these characteristics mean that very small errors in approximating the system's initial state are eventually magnified into

---

74 Thanks to Brian Meloon of the University of Wisconsin for bringing this simplification to my attention.

gross errors about the system's evolution.[75] It is easy to see that if we had only coarse-grained information about a chaotic system's state, and if we were interested in making predictions about the detail of the system's future behaviour, we could not make those predictions over anything but the shortest time scales because of the (normally exponential) expansion of error caused by SIC. This remains true even if we have detailed knowledge of the equations describing the behaviour of the system at a very low level. Thus, such systems are indeterministic *with respect to* information at the coarse level (but this does *not* imply they are indeterministic at the lowest level; i.e., that they are "really" indeterministic—see Chapter 17). But there is another kind of prediction which is unproblematic for these kinds of systems, a type which we should keep in mind as we move on to exploring the framework in which we can represent intelligent systems dynamically.

The kind of prediction which is relatively unproblematic for chaotic systems is possible when we ask a different kind of question about a system's evolution than what the specific detailed phase space location might be. Instead of asking for details of phase space location, we might be interested only in the *attractor* towards which a system is tending. Like all dissipative systems—i.e., systems in which any given volume of phase space shrinks through time[76]—the kinds of chaotic systems we are concerned with have attractors, structures in phase space towards which phase trajectories within a particular neighbourhood of the attractor (called the attractor's *basin*) tend asymptotically. (That is, trajectories which don't begin on the attractor never actually evolve onto it, but they may get arbitrarily near it.) Attractors are invariant under the operation(s) defining the dynamics of the system. They are the large structure analogue of fixed points: an invariant *subspace* rather than a single point.

The interesting thing about attractors in the phase space of chaotic systems is that often they are *strange*; for our purposes, we can simplify the idea of strange attractors by appealing to the property of being infinitely

---

[75] Contra commentators such as Hunt (1987), Stone (1989), early (since corrected!) Smith (1991), Hobbs (1991, 1994), Kellert (1993), and even early Mulhauser (1991), however, this magnification of error needn't completely swamp all attempts at prediction. See also Chapter 17.

[76] It is not impossible for there to be *local* dilation of phase space volumes, even in a highly dissipative system (such as the Belousov-Zhabotinski reaction). But the overall trend must still be for volumes to contract; thus, local dilation in one neighbourhood demands speedier contraction in another.

detailed in a nontrivial way, or *fractal*.[77] Typically, the dynamics of a system on a strange attractor are chaotic, so we still have, for instance, sensitive dependence on initial conditions for trajectories on the attractor itself. This also holds true in the basins of attraction of these strange attractors. Thus, it is altogether possible for a system to be very difficult to predict in the sense we described above, yet for its activity still to be constrained within a particular basin of attraction. Thus, if we are interested only in knowing which attractor a system is near, it can be a very simple matter to know its future state from a measurement of its present state. Even though we may not be able to comment as to the precise phase space location of a phase trajectory's evolution, we can still be sure that the trajectory has not left the basin of attraction in which it started and that in fact it will always get closer and closer to the attractor itself. Under this kind of *complex* coarse-graining of phase space, as distinct from the simple coarse-graining we discussed above as in Figure 23, prediction of the time evolution of a chaotic dynamical system is unproblematic, and chaotic systems are entirely deterministic with respect to this coarse-grained level of description.

With an understanding of dynamical systems and the basics of chaos theory to hand, we now move to an exploration of a dynamical systems framework in which we may outline relationships between high, low, and intermediate level descriptions of intelligent systems. I suggest three different metric spaces representing the level of intentional states of mind, the level of computational relevance, and the level of actual physical state.

13.3 Three Greeks

13.3.1 λ Space

The first space, which we shall call λ, is the most straightforward: it is simply the ordinary physical state space representation for every entity[78] which is functionally relevant to the system we are wanting to represent. The metric for this space is the ordinary euclidean distance measure—the

---

[77] Technically speaking, a fractal is not necessarily strange, and activity on a strange attractor is not necessarily chaotic; but these details do not concern us here.

[78] Here we assume for simplicity a quasi-classical treatment in which Hilbert space wavefunction descriptions of the relevant particles or whatever are entirely decohered. (Chapter 4 and Mulhauser 1995 in press)

square root of the sum of the squares of the differences between two points along each dimension. In practice, we would never represent an intelligent system such as a brain in λ space because of the extraordinary dimensionality involved. We will, however, use it as the basic physical foundation upon which the other two more complex spaces rest.

13.3.2 ω Space

In the second space, which we shall call ω space, we are concerned with "computational relevance": instead of representing individual particles or some such, we represent functionally relevant components of the intelligent system. Thus, to represent a neural network, we might include dimensions for each neuron's output frequency, its level of fatigue and the efficacy of its synaptic connections, plus dimensions describing neuromodulator distributions and other extra-neural factors. The metric for this space can also be something like an ordinary euclidean distance, although we must be aware that the mapping from real state space to ω space will not always be straightforward and that nonuniformities in the mapping may mean that transition times between two ω points may vary because of differences in the distances between the various sets of distinct λ points which may be mapped to the same two ω points. Indeed, the topological transforms mapping manifolds in λ space to those in ω space may be fuzzy, on account of the vague character of components like neurons. (I.e., the precise boundaries indicating which particles should be included in a particular neuron may not be well defined.) We will have more to say about this kind of problem as it relates to the next representational space.

Note that something like ω space, unlike λ space, may serve a practically useful rather than just a theoretical purpose. While it would certainly be difficult to represent the entire human brain in such a space, there is no problem with representing smaller subnetworks directly. For more complex networks, ordinary ω space is still useful as the basis for a coarser-grained representation with dimensions for representing the overall activity of *populations* of neurons (such as columnar arrays in the neocortex) rather than of individual neurons.[79]

---

[79] Having been accused by at least one reviewer of confusing the issue by introducing this space, note that many researchers already either appeal directly to such a dynamical systems representations of neural networks or make a case for the usefulness of such representations. These include Choi and Huberman 1983, Kolen and Pollack 1990, King

13.3.3 ψ Space

The final space accommodates a very high level description of what we might call ψ states, or something like "subjective states of mind", or "intentional states". The parameters defining this space would be largely independent of those describing neural phenomena in ω space. They would include every psychological parameter available to the introspection of a subject, such as descriptions of anxiety, happiness, desire, arousal, fatigue, anger, and so forth. It is at this level where we might describe Horgan and Tienson's (1993) "cognitive state transitions", or changes from one mental state to another.

The dimensions of ψ space may alternatively describe the "contents" of the self model data structures which we explored in the earlier part of this dissertation. In the end, I believe both ways of interpreting ψ space will be of considerable value in understanding relationships between low level dynamics and higher level experience. Because any particular state in ψ space describing data structures will, for a given individual, be correlated with exactly one state in ψ space on the alternative interpretation—and recall from our earlier discussions that this is an empirical correlation, not a logical implication—for the kinds of analyses which follow, either interpretation will do, and for the most part we shall take no great care to distinguish them.

Notice that the usefulness of the space interpreted in these ways stands or falls with the usefulness of the kinds of descriptions common to artificial intelligence. The field of artificial intelligence operates in part on the hope that the relevant aspects of cognitive states may ultimately be described at the level of propositional (symbolic) information. Researchers work on the idea that cognitive processes amount to manipulations of these descriptions which may be quantified in a systematic way and mimicked by computers or other constructed devices. The usefulness of ψ space does not depend on this stronger notion, but it does depend on the legitimacy of something like the former. If the artificial intelligence project is on target in terms of how it describes cognitive states, then the propositions being manipulated serve as the

---

1991, Wilson and Bower 1992, Pollack 1992, Chapeaublondeau 1993, and Horgan and Tienson 1993; not to mention others we will discuss in more detail elsewhere.

basis for phase space representation.[80] If the artificial intelligence project is not on target in this limited respect, it is hard to see what other basis we might appeal to for the quantification of psychological states.

The metric for such a space (on either interpretation) might again be based on the ordinary euclidean distance measure—but we must keep in mind that questions about distance will be phrased in terms of *total* mental states or states of a data structure. Thus, we wouldn't be asking whether something like a mental state including the proposition "I like bran flakes" is closer to one with "My toe hurts" or to one with "Bachelors are unmarried and male". By way of analogy, we might think of one of those personality tests with five hundred questions to be answered something like "agree strongly" or "disagree mildly", except with a continuum for each answer; here we will be asking questions about the distance between two sets of scores represented as points in five hundred dimensional space. The distance may not correspond to anything about which we are accustomed to thinking, but of course that doesn't mean it isn't a legitimate and useful distance for theoretical accounts.

More importantly, perhaps, we must keep in mind that the parameters defining the dimensions of the space when it is understood to represent mental states rather than states of a data structure typically will themselves be vague terms and that it may not make sense to attribute a well defined distance between two distinct points. (It might be very difficult to attribute a precise real number description to a measure of something like happiness, for instance.) And as in the case of transforms mapping λ space to ω space, the topological transforms relating manifolds in λ space to those in ψ space are liable themselves to be fuzzy. Moreover, because subjective states of mind are multiply realisable in neural terms (that is, two distinct ω states may correspond to the very same psychological state), transitions between two neighbourhoods in ψ space may not always take the same amount of time. Whatever the mapping may be from λ state or ω state to psychological state, there will always be the possibility of this difference in transition times. For instance, given

---

[80] Of course, we might work with a superset of the kinds of propositions common to artificial intelligence; few AI researchers, for instance, try to model the time evolution of jealousy or lust. Moreover, adopting the propositional style of description does not commit us to adopting anything resembling the kinds of *operators* AI researchers apply to their propositions, and it does not require a commitment to computable or computationally tractable transitions between psychological states so described be.

any two distinct physical states which map to the same psychological state, one of them may be closer in λ space than the other to another set of physical states which map to a different psychological state. Thus the distance required to traverse the space may also be shorter for the one which is nearer. Multiple realisability, then, suggests that while we might apply a standard euclidean distance to ψ space, we must keep in mind that our answers must be applied to what is at this level a fuzzy and temporally variable reality.

More significantly, multiple realisability also suggests that ψ space may be treated as warped: it is a general Riemannian space in which elliptic, hyperbolic, or neutral geometries might apply according to which neighbourhood we are considering. We can understand this just by considering what kinds of topological transforms might be applied to translate a manifold in ordinary physical λ space into a manifold in ψ space. (The explanation could just as easily be phrased in terms of transforms from ω space to ψ space.) A manifold or a set of disconnected volumes in ordinary physical space might map, for instance, to one single point or perhaps a curve in ψ space. This is just what multiple realisability means. Moreover, the topological transform, or the mapping from λ space to ψ space, is liable to vary in detail according to the neighbourhood in question. That is, in some areas of real state space, large volumes might be mapped to single points, whereas in others small volumes might map to large manifolds. The result of such a mapping, apart from the loss of information which goes with any coarse-graining, no matter how complex, is that there is no guarantee of a consistent geometry across ψ space. We might expect the geometry of ψ space to look something like that of four dimensional spacetime with massive bodies scattered about. But because the warpage of the space is not caused by anything akin to simple masses, which induce a mathematically uniform deformation of spacetime, the geometry of ψ space is liable to be far more complex.

There are a number of observations we can make about the relationship between these three different metric spaces. In the following we will outline a few of these before eventually proceeding in the next chapter to a discussion of one particular argument which has emerged from something like this way of viewing different levels of description of the same intelligent system.

13.4 Space Relations

First, it seems clear that at any given time slice, the "number" of distinct ψ states will be less than the number of both ω states and λ states. Yet, while the "number" of ω states will generally be less than the number of λ states, in particular neighbourhoods of the phase space of chaotic neural networks, the number of ω states may actually be greater than the number of λ states. The first observation comes from the multiple realisability of intentional states and of computationally relevant components such as neurons, while the second is an immediate consequence of SIC. This is a refinement of Pylyshyn's (1984) conjecture that the number of computationally relevant brain states is always less than the number of physically discernible states.

There is also an important observation to be made here about what we mean by "number" of states in a particular space. As Deutsch (1985a) notes, referring to theoretical work by Bekenstein (1981) on the thermodynamics of black holes, any physical system enclosed by a surface with an appropriately defined area A can have at most an extremely large but finite number N(A) of distinguishable access states:

$$N(A) = \exp\left(\frac{Ac^3}{4\hbar G}\right)$$

where c is the speed of light, and the denominator is four times the product of Planck's reduced constant and the gravitational constant. This reveals something important both about the way we must use the dynamic spaces discussed here and about the way almost all mathematical models should be applied to reality.

In particular, if the number of distinguishable access states of a bounded physical system is limited by the equation above to some very large but finite number, then strictly speaking there is only a finite class of points in the state space of any bounded physical system which can have measurable significance to us. But a finite set of points marked off next to each other has zero length. That is, only continuous segments made of an infinite number of points have length. Thus modelling any real physical system must require either a discontinuous space made of discrete points and "empty space" between them (as well as discontinuous dynamics in the space) or, alternatively, a continuous space with fuzzy points

constituting a continuum. In the latter case, the continuous fuzzy regions surrounding each accessible state provide the infinite class of points to allow for overall continuity. This is a general point which is also significant for the discussion later about modelling Nature with an infinitely detailed number system. Fortunately in this particular case we already have a handy way of circumventing the problem; namely, we have already noted that the mappings from λ space to the other two spaces are themselves fuzzy.[81] Thus we can expand ψ space from a space with essentially no volume to a continuous space with positive volumes; the same applies to ω space. We must simply keep in mind when using the dynamical spaces that two points very near each other may simply be indistinguishable: we are operating with models applied to vaguely mapped spaces, and our "answers" must be fuzzified appropriately. This is exactly analogous to our later conclusions about using real number models which include an infinite amount of detail.

Two other observations about this dynamical systems framework follow on from these kinds of points. The first is that whenever we are modelling a system at a high enough level of description that there is some loss of detail from a lower level where dynamics are deterministic, there is at least the *possibility* that dynamics at the higher level will be nondeterministic. If there are "fewer" states at a given level—say, the ψ level—than at a lower deterministic level—say, the λ level—then it follows from what mathematicians call the "pigeon hole principle" that there will be at least one ψ description into which more than one λ description must go. (Note that this is essentially a restatement of the main point of multiple realisability.) Thus, recalling Figure 23, there may be points which are distinct at the lower λ level and which have unique time evolutions at that level but which are the *same* point at the higher level before ultimately tracing different time evolutions at that higher level. (This is independent of whether we are tracking mental states or data structure states in ψ space.)

The second related point is that because of this loss of information as we move to higher levels, the topological transforms relating manifolds at lower levels to higher levels (which, as noted above, might not

---

[81] Strictly speaking, we could also say that the mapping from reality to λ space is fuzzy, because a system represented in λ space has only a finite number of access states distinguishable in λ space .

themselves even be continuous) may not be invertible. That is, once we have transformed a manifold in, say, λ space, to one in ψ space, it may not be possible to get the original surface back by inverting the transform. The easy counterexample to invertibility occurs around a singularity where a volume of points in λ space maps to a single point in ψ space.

Finally, we move on to some brief comments about applying this dynamical framework to existing theories and explore how new theories might be formulated within the schema.

13.5 Technoflash or Good for Something?

13.5.1 Existing Theory

Within this framework, we may formulate succinctly theories about phenomena in intelligent systems which rely on the relationships between different levels of description. For example, one main point of Horgan and Tienson's recent work into the computability and computational tractability of cognitive state transitions can be put very economically in the terms of the new schema. Specifically, they suggest that transitions between ψ states are not always computable solely on the basis of ψ level information, because the same ψ state might supervene on two or more distinct ω states which could ultimately evolve along trajectories distinguishable not only in ω space but also in ψ space. Particularly in chaotic systems, this means the same ψ state might branch into two different ψ states on the basis of lower level sensitive dependence on initial conditions. At the least, information loss may result in ψ state transitions which are computationally intractable even in the case of tractable ω computability. In effect, theirs is a comparison of ψ state overdeterminism with respect to the ω and λ levels and underdeterminism with respect to the ψ level itself. We might extend their ideas by asking about the characteristics of ω level indeterminism on the basis of underlying λ dynamics and chaos in real space.

My own thoughts on applying recursion theory to chaotic analogue systems appear in Chapter 15; while confined primarily to noncomputability at the λ level, they might when paired with this representational schema create a foundation for exploring problems of computability at higher levels of description. The schema might also be a

useful point of departure for examining further ramifications of our refinement of Pylyshyn's conjecture suggested above.

13.5.2 New Frontiers

As for new ways of using the schema for exploring interactions between low level chaotic dynamics and characteristics of behaviour at the introspectively accessible level, we might gain new insights into psychological questions about creativity and problem solving, philosophical questions about free will and the relationship between reasons and causes, and computational questions about implementing artificial intelligences.

For the psychologist, it could be useful to try to describe the relationship between creativity and SIC or other dynamic properties at the ω or λ levels. Because a minor perturbation at a low level might result in a system's taking a new path at a higher level, this property of chaotic systems might provide some clues about sudden flashes of inspiration or "lateral thinking" which, while entirely deterministic at lower levels, may appear at the introspective level as discontinuities in our train of thought. (We return to this idea in the next chapter.)

The speed with which a chaotic system may visit different neighbourhoods of its phase space, coupled with the distributed representational abilities of neural networks, might also offer a partial account of what appears to be content addressable memory. That is, a network visiting large areas of its phase space "looking" for a pattern to match might be one mechanism subserving content addressable memory.[82] Understanding the rôle of content addressable memory may well be critical for making sense of creativity, problem solving, and the so-called frame problem. There is also some evidence (Tsuda 1994b) that low level chaos enhances learning and the actual organisation of memory.

For the philosopher, explorations of interactions between levels might also provide insights into the appearance of free will. Because so much of ω and λ level dynamics is not available to introspective awareness at the ψ level, it is clear that *apparent* contracausal behaviour at the ψ level might be possible on the basis of entirely deterministic λ level

---

[82] A forthcoming paper based on the interlocking ring architecture of Chapter 10 discusses content addressable memory in more detail. Another forthcoming paper engages the frame problem more directly from a neural network perspective.

causation. Yet this apparent contracausal behaviour needn't be without precursor *reasons* at the ψ level.

Consider the evolution of a single intelligent system traced at both the ψ level and, say, the λ level. The sorts of things we track at the higher level are measures of psychological parameters or, alternatively, descriptions of self model data structures, and states in this space, while they may be related to each other by stochastic generalisations describing which are likely to follow which, they needn't have strict *causal* connections between them as such. (To pick an example out of the air, fear of heights might lead to an amorous admiration of a Russian trapeze artist, but then again it might not, and it certainly needn't *cause* it in the strictest sense of the word.) Connections between λ states, however, generally *are* causal. Yet the phase trajectory in λ space, described causally, and the phase trajectory in ψ space, described in terms of psychological generalisations or *reasons*, describe the evolution of the very same intelligent system. Just as it is a category error to conflate reasons and causes, on this view it is also a mistaken use of levels of description. But we can see here that the two may be brought together, without reducing one to the other or committing a category error. Any sequence of actions would have traceable reasons in ψ space but causes only in ω or λ space.[83]

A full treatment of questions about reason and causation must await another occasion. But there appears to be considerable philosophical mileage in the idea that as far as we can be introspectively aware, our behaviour is governed by generalisations linking types of ψ states, generalisations for which exceptions often occur, but that our behaviour is still entirely determined at lower levels in a sort of "introspectively invisible" fashion. As far as the ψ level is concerned, this account is not too different from the way it looks when we introspectively consider our own behaviour. That is, it *seems* to us introspectively (or to me, anyway!) that we do have particular general patterns of behaviour but that we can always violate those patterns *when it suits us* (not when we're struck by a sudden fit of indeterministic irrationality). Deterministic but chaotic low level dynamics might provide both a possible account of the elusive "when it suits us" as well as a way of reconciling causal determinism with the appearance of reasoned (not necessarily contracausal) free will. I

---

[83] See Mulhauser (1993e) for an exploration of this idea in the form of a psychological Principle of Sufficient Reason.

believe this rough sketch of an approach to the problem of free will complements rather nicely Dennett's (1984) compatibilist position.

Finally, the quest for artificially intelligent systems could be aided by an exploration of the question of what cognitive state transitions at the ψ level are served by sensitively dependent processes at lower levels, or at least an exploration of what ψ level transitions are nondeterministic at that level but are deterministic at a lower level. Gaining partial answers to this question would tell us what processes can be modelled at higher levels with lower degrees of detail and which must be implemented at lower levels with more careful attention to detail.

For instance, few would dispute that an artificially intelligent system needn't model every single behavioural property of every single neuron in a given human's brain in order to display some of the cognitive faculties of the human. Yet it seems nearly as obvious that for many cognitive faculties we might wish to model, attempting to mimic only the highest level behaviour of the human would be unworkable (either indeterministically unpredictable or computationally intractable) because of underdeterminism at that level. On the face of it, there should be some boundary area between the two extremes which would allow an artificial system to mimic an acceptable degree of the human's subtlety without unduly burdening computational resources by modelling unnecessary details. Until a better understanding of inter-level dynamical relationships is achieved, any choice for this boundary must be at least partly arbitrary.

In short, I believe questions about the relationship between the introspectively available level of description or the level of description of the self model and the finer, lower levels of description are important ones. The framework I have suggested is substantially "underdetermined" itself, in that it may not yet be fleshed out with enough detail to allow the formulation of any but the most rudimentary theories or observations. But it is a possible starting point for one approach to understanding the mind-brain as a rich dynamical system.

14

Determinism and the Topology of Mind

In the manuscript of his recent Cambridge lecture series on chaos (Smith 1993) and in a commentary at the July 1993 Conference of the European Society for Philosophy and Psychology, Peter Smith has strongly criticised one of the arguments which has emerged from the preceding way of viewing the relationship between psychological states and the underlying neural activity.[84] In particular, Smith has criticised the notion that any conclusions concerning so-called anomalous monism can be drawn from the sensitive dependence on initial conditions of chaotic neural subsystems of the brain. The problem is related to the various conclusions which might be drawn about deterministic or indeterministic evolution at the level of subjective experience depending on how we conceive the mapping of sets of points in λ space to ψ states; it is important to our overall project to understand how this relationship between material substrate and subjective experience may work.

14.1 Mind, Temperature, and a Game of Cards

We begin with the argument Smith attributes to me and claims to be invalid:[85]

1. A given initial type of psychological state can be realised in a variety of physical ways.

---

[84] The material in this section is drawn with minor revisions from my response to Smith's critique in Mulhauser (1993f).

[85] Curiously, the paper which was the subject of Smith's commentary (Mulhauser 1993c) does not include any explicit argument of this kind, although I assume something similar to it in my mention of Horgan and Tienson's (1993) work. The purpose of my own paper was simply to provide a framework for analysing the interactions between various levels of description of the same intelligent system.

2. But the time-evolution of the physical states in question is sensitively dependent on initial conditions—i.e., we may get markedly different *physical* upshots arising from very similar initial states.
3. Hence we can get (with significant probability) markedly different *psychological* upshots arising from the same initial psychological state.

The argument, he claims, is invalid because it is exactly analogous to the following, which is obviously invalid because it has true premises and a false conclusion:

1.* A given initial type of thermodynamical state (e.g. a certain temperature) can be realised in a variety of physical states characterised by different position/momenta distributions of the particles in a gas.
2.* But the time-evolution of a state with a given position/momenta distribution is sensitively dependent on initial conditions—i.e., we may get markedly different distributional upshots arising from very similar initial states.
3.* Hence we can get (with significant probability) markedly different thermodynamical upshots arising from the same initial thermodynamical state.

Smith claims this analogy establishes that this general form of argument is invalid and that nothing about indeterminism at higher levels of description can be inferred from the presence of low level chaos. But compare the following argument, which doesn't involve chaos at all:

1.** A given type of poker hand (such as one pair or a full house or a royal flush) can be realised by a variety of physical card distributions.
2.** But the time-evolution of physical card distributions (given an appropriate algorithm for discarding cards) is "sensitively dependent" on the initial conditions of the cards in the hand—i.e., we may get markedly different physical card distributions arising from very similar initial distributions.

3.** Hence we can get (with significant probability) markedly different final types of poker hands arising from the same initial type of poker hand.

In an earlier IPPE-distributed draft of a document including this chapter, I asserted without argument that the poker argument—apart from simply having true premises and a true conclusion—is in fact valid. Peter Smith has maintained privately that it must be invalid because the general form of argument is invalid, and other readers have expressed confusion over the point. So I feel compelled to make a brief diversion to establish clearly that the third argument above is valid, and this diversion will hopefully serve as a point of departure for understanding what is so odd about the relationship between the three different arguments. (In the end, we will find this argument which is at the centre of confusion really isn't useful, but it is helpful for us to sort it all out in order to understand why.) First, just to be sure, let's notice that if the negation of the conclusion of an argument can be shown to entail the negation of any of the premises (i.e., if (~ C → ~ P1) or (~ C → ~ P2)), then the argument is valid. That is, if the negation of the conclusion entails the negation of any of the premises, then it entails the negation of the conjunction of the premises, and by modus tolens the conjunction of the premises then entails the conclusion (because (~ P1 v ~ P2) ≡ ~ (P1 & P2) and if (~ C → (~ P1 v ~ P2)) then clearly (~ ~ (P1 & P2) → ~ ~ C) and (P1 & P2) → C).

But if we negate a simplified[86] 3.** from above, we get:

~3.** It is false that we can get different final types of poker hands arising from the same initial type of poker hand.

And a few moments' reflection reveals that if we keep constant the rest of the background premises about the game's rules and the distribution of the cards in the deck and so forth, then the only way this can be true is if *either* there is only one way to get each type of poker hand *or* the evolution of a hand in the course of the game is independent of what other cards are dealt. The first of these disjuncts is nothing more than the negation of 1.** above, and the second straightforwardly entails the negation of 2.** above. Thus, the negation of the conclusion entails the

86 We can ignore the "with significant probability" as well as the "marked" for the moment because these depend on the *degree* of SIC; we are concerned only with what is left over above.

negation of at least one of the premises, and according to the observation above, the conjunction of the premises entails the conclusion. But since this argument certainly *looks* to be invalid by Smith's same analogy to the thermodynamical argument, what is wrong?

What makes the thermodynamical argument invalid, the poker argument valid, and the original psychological state argument open to debate concerns our definitions of, and background assumptions about, temperature, poker hand, and psychological state, respectively. The thermodynamical argument is invalid because of the hidden premise which maps position/momenta distributions to temperature: temperature *just is* the mean kinetic energy of the particles in the gas (as read off from the position/momenta distribution). No change in the position or momentum of any particle makes any difference to the temperature as long as the mean is the same. This is the simplest of the three arguments because the only hidden premises establishing connections between temperature and mean kinetic energy are definitional.

The poker argument is valid because of the hidden premises which map card distributions to hands: a full house, for instance, *just is* one pair and three of a kind. Changes in cards *do* change the hand when they bring about a change in the types or sizes of relevant sets that can be formed from the cards. Thus given two distinct card distributions which both have one pair, for instance, discarding the three other cards in both hands and adding identical cards to the hands can yield completely different hands, such as a full house in one and two pair in another. See Figure 24 for this example—it is a straightforward case of divergence in the style of Figure 23, where loss of information at a higher level of description (in this case, loss of information about what *specific* cards are making up a hand) enables lower level factors to drive apart states which are identical at that higher level of description. This argument is as easily seen to be valid as the thermodynamical argument is seen to be invalid: just as we already know how to construe temperature, we know what makes a full house.

The psychological argument is open to debate because there are hidden premises concerned with how we are to understand the mapping from physical states to psychological states, and these hidden premises are controversial. (Smith would apparently have us ignore the importance of

the hidden premises in dubbing the original argument invalid, yet I would maintain that even *understanding* what the original argument *means* requires taking on *some* hidden premise about mappings between physical states and psychological states.

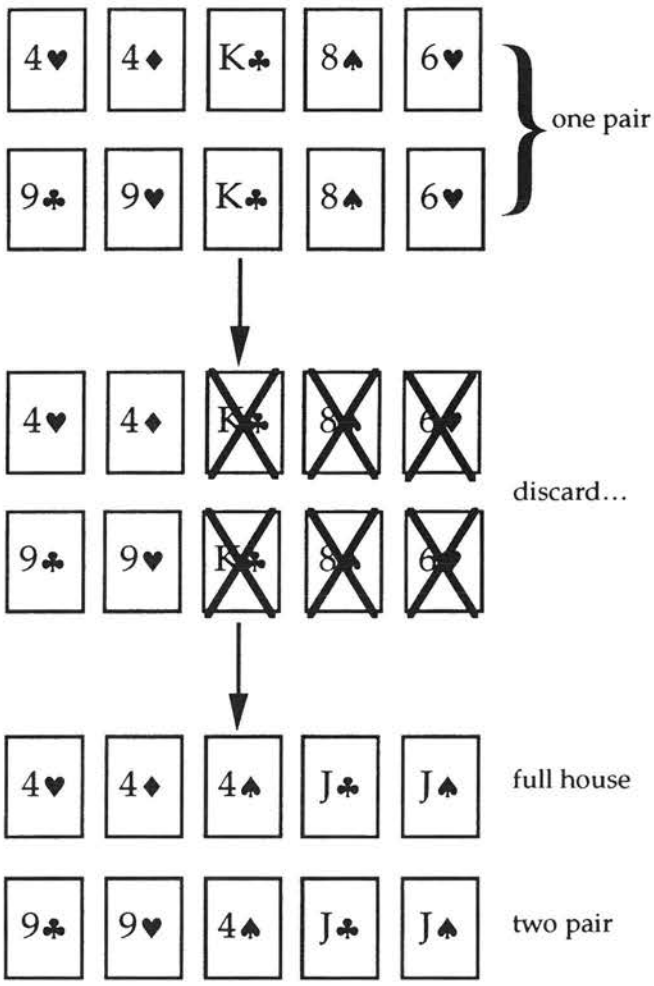


Fig. 24 Sensitive dependence in a card game

Analysing the other two arguments relies on an appeal to hidden premises just as in the present case, but in these cases those premises—concerning how to construe temperature and how to construe a poker hand—are uncontroversial. (Although, it is still fair to say that all three of these are bad arguments in the sense that too much crucial information is left in the hidden premises rather than being made explicit in the arguments themselves.) Now we must examine some of the ways we might construe this mapping from λ space to ψ space.

14.2 Mind Mapping—Finding Our Way

With respect to what sort of mapping scheme we *should* adopt, Smith wonders first if perhaps I have fallen prey to the naïve assumption that psychological states are simply a uniform coarse graining of physical states and then goes on to suggest, loosely following the more recent line of Freeman and colleagues (for entry into the complete literature originating from the research , see Freeman 1964, 1972, 1975, 1979, 1987a, 1987b, 1988, 1989, 1991a, 1991b; Freeman and Skarda 1985; Skarda and Freeman 1987; Yao and Freeman 1990; Eeckman and Freeman 1991), that psychological states correspond to large structures in phase space which might be strange attractors. Fortunately I haven't fallen prey to the uniform coarse-graining assumption, as should be apparent from my comment in the paper under debate that psychological state space should properly be treated as a general Riemannian space. The reason ψ space should be treated as a general Riemannian space was simply that there might be complex mappings from λ space to ψ space which varied across the space as a whole. This notion is incompatible with a simple uniform coarse graining, which would yield a uniform geometry across the whole space. But let's examine the idea that psychological states correspond to states near strange attractors. Smith provides no indication of how we are to construe the word "near" here, and in private communication he has objected to my simplifying it to anything other than states within some appropriate ε of an attractor. But lacking for the moment any hint of how to determine this ε or any clear reason not to simplify the idea further, for the present purposes I believe it introduces no confusion simply to replace "lying near a strange attractor" with "lying within the basin of attraction of a strange attractor". In making this simplification we are thus not always addressing precisely what Smith has intended, but our discussion remains useful for our own purposes.

Smith notes correctly that if a given psychological state corresponds to a physical state lying near a strange attractor, then we may observe divergent physical phase trajectories from arbitrarily similar initial conditions while retaining the same psychological state (because those phase trajectories, while divergent, remain near the same attractor). This provides what he calls "micro-chaos but macro-psychological stability", and, he continues, "the move from one dynamical state (defined by its

attractor) to another as control parameters change can in fact be as deterministic as you like" (Smith 1993, p. 76). Such a picture allows for the wild low level behaviour characteristic of chaotic systems without making the high level psychological behaviour similarly wild.

It is worth noticing, incidentally, that arguments over macro-level stability in the brain—which Smith apparently assumes is an obvious feature—are subtle and by no means conclusive. For instance, Wright and colleagues (Wright, et al 1993) have suggested a model for chaotic EEG data which effectively ties low level cortical chaos to stable group dynamics, but their work is vulnerable to a number of criticisms stemming from their failure to take account of poorly understood features of cell dynamics (Goertzel 1994) and their reliance (Wright 1990) on noise injected via the reticular formation. (Tsuda 1994a, drawing on Kaneko 1990) While they have partly defended themselves (Wright, et al 1994), the matter remains highly contentious.

But returning to Smith's idea, I believe he *has* correctly sussed out a description of something like what must happen in response to changes in parameters such as ion concentrations and neuropeptide distributions. Such physical changes correspond to the control parameters to which Smith refers. Some neuromodulators, for instance, influence the way output frequencies of whole cell families vary in response to afferent signals. Thus they change in a global way the shape of the network's possible trajectories through phase space, and they may well lead to just the kind of change in psychological state which we would expect under Smith's picture. But to accept this picture as the whole story is to ignore altogether the rôle of perturbations (corresponding to changes in afferent signals) in shifting the state of a neural network from the basin of one attractor to that of a second attractor, coexistent with the first.[87]

---

[87] Smith has objected in private communication that he intended "control parameter" in a more general sense, to include changes in afferent signals. But the standard use of "control parameter" refers to either a constant or a coefficient of one of the terms of the equations describing the dynamics of the system. It does not generally refer to anything like what we mean here by a change in afferent signals, which corresponds simply to a possibly abrupt change in the system's location in phase space but *not* to any overall change in dynamics. In any case, if we allow Smith's broader notion of "control parameter" as a term which does include simple changes in the position of the system in phase space, then the notion that psychological state changes in response to changes in afferent signals are deterministic with respect to psychological state level information becomes highly dubious. As we shall see, it is trivial to show that under a simple mapping of physical states to psychological states in terms of attractor basins, psychological state response to a change in afferent

It is an elementary observation about even the simplest artificial neural networks (and thus presumably it applies to the far more complex biological neural nets) that the state spaces[88] of such networks include multiple attractors. In the lingo of artificial neural nets, this is just another way of saying that the output patterns vary according to changes in the input patterns. To use a hideously simplified discrete example, consider any pattern recognition network which fires a single output unit in response to each of ten possible input patterns (the technical details are irrelevant). Then we can put the point very crudely by saying that all the states of the network corresponding to the presence of an input pattern which looks like a "1" will be attracted to a manifold in state space where the "1" output unit fires and the others don't, all the states corresponding to the presence of an input pattern which looks like an "8" will be attracted to a manifold in state space where the "8" output unit fires the others don't, and so on. All ten attractors are coexistent in the state space of the network, and the state of the network evolves in response to changes in the afferent signals, *not* in response to changes in global control parameters.[89] Smith's picture apparently ignores altogether the response of a neural network to such changes in afferent signals.

14.3 Indeterminism and Topology

If we accept the hypothesis that psychological states correspond to basins of attraction (or Smith's hypothesis that they correspond to areas within some ε of an attractor), the existence of multiple attractors in the state space of a neural network simply means that cognitive states can change in response to changes in neural input. Indeed, it seems almost obvious that this is at least as large a piece of the puzzle as Smith's changes in response to modifications of the control parameters: my psychological state changes when presented with a red apple *because of* the change in visual input, my psychological state changes when the orchestra begins to

---

signals can be entirely nondeterministic with respect to complete information about the signal change and about the original psychological state.

[88] Here we mean the state space describing the firing patterns of nodes in response to inputs *after* training; of course it is also useful (and more frequently actually used) to describe the evolution of a net through a state space of connection strengths *during* training.

[89] Of course, we *could* make changes in such control parameters, and the shape of the possible phase trajectories would be altered accordingly; the point is that there are other ways the "psychological state" of the network could change.

play *because of* the change in auditory input, etc. It is not that my neuropeptides aren't playing a rôle, but so are changes in the output frequencies of cells responding to environmental input.

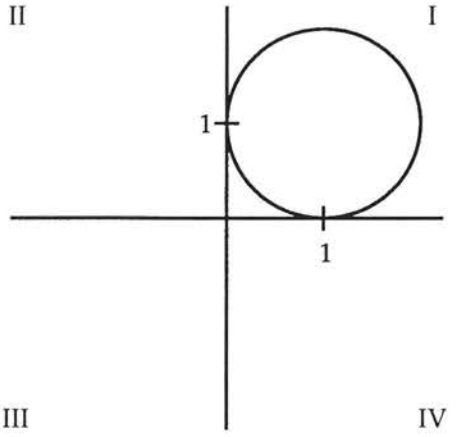


Fig. 25 A bounded space of "identical" mental states

It is worth noting that once we accept the presence of multiple attractors in the phase space of a neural system, we needn't even appeal to chaos to establish something like psychological state indeterminism.[90] We can illustrate with a trivially simple example. Consider a two-dimensional state space with four attractors such that the basins of attraction are marked off by the four quadrants of the cartesian coordinate system. Now suppose we are given an open circle of radius one, tangent to both the horizontal and vertical axes in quadrant I. This possibility is illustrated in Figure 25. Since all the states within this circle are in the same basin of attraction, they all map to the same psychological state.

Now we perturb the system (i.e., make a change in inputs) corresponding to a nudge parallel to the unit vector $\left\langle \frac{-\sqrt{2}}{2}, \frac{-\sqrt{2}}{2} \right\rangle$. In an artificially simple case, with a vector of magnitude $\sqrt{2}$ and ignoring the flow within the basins of attraction, the circle is now centred on the origin. (Depending on the character of the flow in each of the four basins of

---

[90] Smith has observed in private communication that we shouldn't really be talking about *anomalous monism* here, since the original Davidsonian meaning of that term referred to physical monism without strict psycho-physical laws. Understanding psychological states as supervening upon volumes of points near particular attractors or within their boundaries of attraction is compatible with and even suggests that the correlation between ψ states and λ states is law-like, the *denial* of anomalous monism.

attraction, the circle may of course be deformed, and it may be necessary to apply a slightly different nudge, but the details are unimportant; all that matters is that we have now moved some portion of the circle over the attractor basin boundaries.) This is illustrated in Figure 26. We can easily see that a physical nudge to the same psychological state can give rise to four different possible psychological states, according to the differences in underlying physical states.

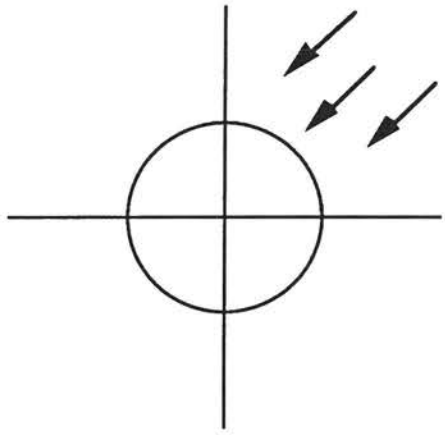


Fig. 26 "Identical" states after a nudge

Well, then, have we dispensed with the need for chaos altogether? We can see from the example that all that is required to observe high level indeterminism on top of low level determinism is a nudge that shifts any set of arbitrarily similar states lying near one attractor to a new neighbourhood where a basin boundary will intersect the translated set. These nudges, to use the phrase Smith applies to changes in control parameters, "can be as deterministic as you like", but that doesn't mean the subsequent evolution, described at a high level, is deterministic.[91] The point is that even with respect to complete information about the change in afferent signals (i.e., the environmental interaction) and complete information about the original psychological state of a cognitive system, psychological state evolution (i.e., response to the environment) may be entirely nondeterministic. And, most significantly, the nondeterminism does *not* come from the environment—since we are offering complete

91 This is why I noted previously that on a broad construal of Smith's "control parameters", the idea about deterministic changes in psychological state is dubious, and in fact it is incorrect.

information about the environmental interaction as a given—but from the *response* of the cognitive system as described at the psychological level. But this isn't chaos, this is elementary dynamics.

It's true: it appears that something like psychological state indeterminism follows immediately unless we are prepared to ignore basic neurophysiology and arbitrarily restrict the attractor changes that count to those which are brought about by changes in global parameters describing ion concentrations, neuropeptide distributions, and the like. The rôle here for chaotic dynamics, and for their associated fractal basin boundaries and sensitive dependence on initial conditions, is to make this psychological level indeterminism more interesting.

Let's return to the argument Smith attributes to me. The gist of it is that sensitive dependence on initial conditions in the low level dynamics of a system can magnify tiny, apparently psychologically irrelevant changes in physical state at one time, into physical changes which are psychologically significant later on. If the picture of strange attractor evolution, or cognitive state transition—the picture obscured in the argument's hidden premises—offered by Smith were complete, this conclusion about sensitive dependence on initial conditions would be false. But when we consider the importance of afferent signals in the way real neural networks function, this conclusion about sensitive dependence on initial conditions becomes true.[92] The reason is that a tiny change in physical state, insufficient to lead to a new psychological state all by itself, *can* cause the system to enter a different psychological state than it otherwise would have the next time it is "bumped" by a change in afferent signals.

To illustrate, consider two arbitrarily similar states within the circle in Figure 25. Perhaps after some time, maybe under the influence of SIC, the states have evolved into the upper left and the lower right of the circle. They are still the same psychological state, because they are both in quadrant I, but when we make a change in afferent signals corresponding to the nudge illustrated in Figure 26, one trajectory now lies in quadrant II and the other in quadrant IV. The presence of sensitive dependence on initial conditions is neither a sufficient nor a necessary condition for

---

92 The truth of the conclusion rests on the hidden information about the importance of afferent signals and *not* on SIC, however; the hidden information reveals why the argument is rather uninteresting.

cognitive transition indeterminism, but when present it makes that indeterminism more interesting by increasing the sensitivity of the system to afferent signal "bumps" and by making that increase in sensitivity even more unpredictable with respect to ψ level information.

As a final note, it is interesting to observe that characteristics of the dynamical flow at the λ level or the ω level may determine the kinds of probabilistic predictions which can be made about ψ level responses to environmental interactions. For instance, since one "bump" in the above example is sufficient to determine in which of the four portions of the circle in Figure 25 the system began, under a fairly uniform and "slow" ω level flow, we might be able to predict that a nudge of a given magnitude in another direction would be highly likely to land the system in a psychological state corresponding to where the original quarter-circle volume of phase space would land under two consecutive nudges. But under, say, a "vigorous" chaotic flow at the ω level, the post-nudge trajectory may have wandered very rapidly into a much larger area, making such subsequent predictions impossible. The situation is faintly reminiscent of the obliteration of information about orthogonal observables which occurs with the measurement of a maximal quantum observable.

14.4 Back to the Mapping—Where Were We Going?

To recap, then, we've established first that the apparent analogy between Smith's thermodynamical argument and the one he attributes to me is insufficient to show the latter invalid. Interpreting the three different arguments depends on a number of hidden premises. In the case of the original psychological argument, we've seen that making sense of it depends on how we construe the mappings from physical states to psychological states. If we adopt the suggestion that psychological states correspond to classes of trajectories near particular attractors, we gain a useful picture of how psychological states may change in response to modifications in global parameters. But when we add in the realisation that networks may move to new attractors in response to changes in afferent signals, we gain two further useful conclusions. First, something like psychological indeterminism follows immediately, without recourse to chaotic dynamics. Second, if the system is in fact chaotic at low levels,

we gain a richer variety of cognitive transition indeterminism. Thus Smith was entirely correct in objecting that appeals specifically to chaos do not add argumentative force to the basic notion that dynamics described at the psychological state level may be indeterministic although they supervene on deterministic dynamics at a lower level. At the same time, along the lines of the original paper, an understanding of low level chaotic dynamics is helpful for understanding what are plausible accounts of the mapping from physical state to psychological state and for understanding subtle properties of high level cognitive transitions and the kinds of predictions about them which are possible; a representational schema for speaking about the relationship between dynamics at various levels of description is useful.

Next we will consider low level chaos itself and conclusions about it which are largely independent of how we may relate low level dynamics to higher level properties of intelligent systems yet are significant for how we might understand the possible ways of modelling an intelligent system such as an instantiated self model data structure.

15

Computability and Analogue Chaos

We have seen some very basic ways in which chaotic behaviour in neural networks might be relevant to the self model data structures they may instantiate. In what follows, we explore a more technical aspect of chaos in neural networks. Specifically, our concern will be with what relevance, if any, the analogue nature of real biological networks has for simulating their behaviour on digital computers. This is a significant relationship to understand, because if analogue networks can be simulated satisfactorily by digital computers, and if we believe human self models are instantiated by analogue networks, then human style self models can also (barring any other problems) be satisfactorily simulated by digital computers. Human cognition in this case reduces to little more than a very sophisticated Universal Turing Machine program exploiting some very sophisticated input and output channels. Indeed, this is the fundamental assumption of the strong AI project.

On the other hand, analogue neural networks might have special properties which make them inherently unsuited to digital simulation. Of course it is possible that just as digital computers can themselves be perfectly simulated at the computational level despite their probabilistic semiconductors, so, too, might self models be perfectly amenable to simulation at the data structure level in spite of any special characteristics of their neural instantiation. But finding "special characteristics" at least establishes the *possibility* that self model data structures instantiated by analogue neural networks may not be simulated digitally. Our assumption is that the existence of self model dynamics which characteristically cannot be simulated satisfactorily implies instantiating wetware which also cannot be simulated satisfactorily. Here, our strategy is to affirm the consequent of that implication; thus, we won't *prove* that any dynamics at the level of the self model are inherently difficult to

simulate, but we will block the corresponding inference from the denial of the consequent to the affirmation that human style self models *can* be simulated digitally without hindrance.

In what follows, we examine the assumption that the computability of a function governing the behaviour of a nonlinear analogue dynamical system guarantees that that dynamical system may be satisfactorily simulated by digital means. After a brief overview of recursion theory, we consider two observations about chaotic analogue systems which suggest some aspects of their behaviour may not be captured by digital simulations. We conclude by placing these observations in a broader recursion theoretic context and by commenting on their potential relevance to chaotic analogue subsystems of the human brain.[93] Later in this dissertation (see Chapter 17) we shall examine a recent development which apparently offers real evidence of the kinds of difficulties explored more abstractly in the following discussion.

15.1 Doing It and Doing It For Real

Mathematicians usually assume that the computability of a function governing the behaviour of a dynamical system is both a necessary and a sufficient condition for its being possible—given appropriate resources—to simulate effectively that dynamical system with a digital computer. In other words, the assumption is that it is impossible for the real dynamical system to behave in such a way that a digital simulation could not behave arbitrarily similarly, to within a degree of accuracy limited only by the computational resources available.

This is not an unreasonable assumption. The requirements for a function's computability—sequential computability and effective uniform continuity—appear, on the face of it, to preclude any kind of behaviour at all in the dynamical system itself which cannot simply be read off from an appropriate application of the function to a set of specified initial conditions. Even in cases where necessary errors in specifying the initial conditions of a system mean the time evolution of a real system cannot be predicted precisely, it appears that no phase trajectory which the real

---

[93] Much of the material in this section is reprinted or adapted, with some modifications, from Mulhauser (1993d—please see the Appendix for reprint information) and the early Mulhauser (1992).

system might trace could be qualitatively different from phase trajectories described by simulated evolution.

In what follows, we explore whether this holds true for chaotic systems which are analogue and whose behaviour we thus might analyse in terms of values on the continuous real line. We begin with a brief overview of fundamental definitions of recursion theory. With these definitions and the earlier definitions of chaos in hand, we observe two general properties of chaos over the real line. Finally, we locate these observations in a broader recursion theoretic context and suggest ways in which they might be relevant to the analysis of chaotic areas of the human brain.

15.2 Recursion Theory

At the heart of the theory of computability is the recursively enumerable set. Put simply, any set of natural numbers which can be generated algorithmically—alternatively, by a Turing machine—is called recursively enumerable. The class of all such algorithms, or Turing programs, generates the class of all recursively enumerable sets. For some of these, called recursively enumerable nonrecursive sets, there is no algorithmic test for set membership. It is just such a *noncomputable* set, one whose *complement* is not recursively enumerable, which is the key to understanding noncomputability for a function, number, or sequence of numbers. That such sets exist was first established by Kleene (1952; see also Rogers 1967).

A real number is computable if there is a computable sequence of rationals which converges effectively to it. (Pour-El and Richards 1989) More formally, a sequence of rational numbers $\{r_k\}$ is computable if there exist three recursive—that is, algorithmic—functions over the naturals *a*, *b*, *s*: $\mathbb{N} \rightarrow \mathbb{N}$ such that $\forall k$,

$$b(k) \neq 0 \qquad \text{and} \qquad r_k = (-1)^{s(k)} \frac{a(k)}{b(k)}$$

Such a sequence converges effectively to a real number x if there exists a recursive function over the naturals *e*: $\mathbb{N} \rightarrow \mathbb{N}$ such that $\forall N \in \mathbb{N}$:

$$k \geq e(N) \qquad \text{implies} \qquad |r_k - x| \leq 2^{-N}$$

More intuitively, a real number is computable if it can be effectively approximated to an arbitrary degree of accuracy by an algorithmic method.

For instance, π is a computable number because the successive digits of its decimal expansion can be generated to an arbitrary degree of precision by an algorithm specified in advance. (Note, however, that the question of whether a particular sequence of digits occurs in the expansion of π cannot, in general, be decided algorithmically; the best we could do would be to keep generating the decimal expansion and testing for the sequence—unless and until it did appear, we could not answer the question of whether it might not still appear.) It is significant that "most" real numbers are in fact noncomputable. (Minsky 1967) It is easy to see why: every real number is either computable or noncomputable, but the set of computable reals is only *countably,* or *denumerably,* infinite, while the set of all reals is *uncountably* infinite. This point plays a central rôle in the observations we shortly will make about chaotic analogue systems.

Computability for a *function* was first formulated over three decades ago (Grzegorczyk 1955, 1957; Lacombe 1955a, 1955b), and it requires both sequential computability and effective uniform continuity. Consider a function f defined on a closed bounded rectangle I^q in $\mathbb{R}^q$, where

$$I^q = \{a_i \leq x_i \leq b_i, 1 \leq i \leq q\},$$

a_i, b_i (the "corners") are computable real numbers, and $\mathbb{R}^q$ represents q-dimensional real space. The first criterion is met when the function maps computable sequences to computable sequences: f maps every computable sequence of points $x_k \in I^q$ into a computable sequence $\{f(x_k)\}$ of real numbers. The second condition is fulfilled when a certain algorithmic relationship exists between the euclidean distance separating points in the domain of the function and the distance between corresponding points in the range. Specifically, the condition is met when there exists a recursive function d: $\mathbb{N} \rightarrow \mathbb{N}$ such that $\forall x, y \in I^q$ and $\forall N \in \mathbb{N}$:

$$|x - y| \leq \frac{1}{d(N)} \qquad \text{implies} \qquad |f(x) - f(y)| \leq 2^{-N}$$

Having outlined the technical meaning of 'computability' on which the rest of our points rely, we turn now to an analysis of computability for chaotic systems which are analogue.

15.3 Analogue Chaos

The present observations rely upon the assumption that the continuous real line—as opposed to, say, the constructive rationals—represents the best mathematical framework in which to analyse the

behaviour of analogue systems. While both sets of numbers provide continuity, the reals are an intuitively more "complete" transfinite set and offer a starting point which, on the face of it, is certainly no less plausible than the alternative. Although we will return to a closely related question later when we discuss the applicability of real number mathematical models to a physical world with apparently limited detail, an extended discussion of the merits of each set of numbers for analysing analogue systems is best left to a paper dedicated to philosophy of maths or philosophy of logic. In the present context, we concentrate entirely on the consequences of applying the real numbers.

As we noted previously, there is a nondenumerable infinity of noncomputable numbers on the real line and a denumerable infinity of computable numbers. As Minsky (1967) put it, "most" real numbers are noncomputable. It is often said of typical chaotic systems that their dense set of periodic points includes fixed points of all possible periods. This is a tricky proposition, because *all possible periods* means there is a period for every point on whatever interval of the real line the dynamics of the system permit to be reached. Chaotic systems can be said to have *at least* a countably infinite set of periodic points, and in the context of iterative systems, there can be *only* a countably infinite set of periodic points and a corresponding set of periods. But an analogue system is not subject to this constraint, and there is no reason *prima facie* why there could not be *analogue* systems in nature which truly have fixed points of *all* possible periods. This is no more mysterious than the observation that a discrete model of a rolling wheel limits the wheel's possible rotation angles to a countably infinite set, while an analogue wheel may actually roll through a continuous range of rotation angles from 0 to 2π describing an uncountably infinite set. But then such chaotic analogue systems may have an uncountably infinite number of possible periods but only a countably infinite number of computable points in phase space at a given time or over a given Poincaré section.

If this is true, that there is an uncountably infinite set of possible periods but only a countably infinite set of computable points in phase space (just as there is an uncountably infinite set of possible rotation angles but only a countably infinite set of computable rotation angles), then it is a simple observation that there must be an uncountably infinite number of fixed points with unique periods. Equivalently, there is an

uncountably infinite set of fixed points with noncomputable periods. To use Minsky's term again, "most" phase trajectories, then, have noncomputable periods.

As a followup, it is interesting to note that a point in phase space defined by computable coordinates could have a noncomputable period; likewise, a point defined by noncomputable coordinates could have a computable period. But if *either* the coordinates of a point in phase space are noncomputable *or* the period of a point is noncomputable, then the phase trajectory on which such a point lies is noncomputable. Thus, unless there is some strange reason according to which computable points in phase space musn't have noncomputable periods, not even the full countably infinite set of computable points in phase space will lie on computable phase trajectories (although, of course, the set of computable points in phase space which do lie on computable phase trajectories might well still be countably infinite!).

The second observation about chaotic analogue systems is closely related to this first. If a system is described by a computable function, and if a phase trajectory passes through a computable point at *any* computable moment in time, then it can *always* be located by computable coordinates at computable time values. The reasoning is an obvious *reductio ad absurdum*: if a phase trajectory at a computable temporal and physical location were to evolve into a noncomputable point at some future (or past) computable time, then the computability of the function governing the system would give us a computable method of calculating a noncomputable value.

Now, so far these observations appear irrelevant to the matter at hand: the second applies to any and all analogue dynamical systems, while the first applies to any system with a dense covering of periodic points of all possible periods. So far, we have not exploited the properties of chaos, and these observations seem to offer no problems for effective simulation of analogue dynamical systems in general. But what makes them interesting is their application to systems which are specifically analogue *and* chaotic.

In the case of the first observation and its followup, what is most interesting is that if we consider the class of chaotic systems in which points in the dense covering of periodic points are unstable,[94] all

[94] Recall that in some systems the entire set of periodic points is unstable.

computable phase trajectories diverge from all noncomputable phase trajectories. In an ordinary linear analogue system, this wouldn't be particularly significant, because even though none of the points in a neighbourhood of a noncomputable point would lie on phase trajectories which effectively converged to that of the noncomputable point, it wouldn't be the case that all of them actually *diverged* to the extent that is evident from SIC. Moreover, in a non-chaotic system, we might bound the period of a noncomputable phase trajectory with the periods of neighbouring phase trajectories, such that even without an effective approximation, we could still state an error bound less demanding than that of effective convergence. In a chaotic system, we are guaranteed no such luxury.

Notice, incidentally, that without violating the Shadowing Theorem (see [P*] of Chapter 17), there doesn't seem to be any *a priori* reason why there couldn't be arbitrarily large variation in the periods of two neighbouring fixed points an arbitrary distance apart. Consider two arbitrarily proximal points and their subsequent time evolution through periodic Poincaré sections (i.e., a return map). The two points may remain arbitrarily near to each other through all the Poincaré sections (thus not violating the Shadowing Theorem), yet one may "hit" its original location after, say, five slices (thus establishing its period as five times the interval between sections) and the other after, say, five billion slices.

Thus, the first observation shows that not only may there be an uncountably infinite set of unique phase trajectories in the phase space of the relevant type of chaotic analogue system which are noncomputable and cannot be effectively approximated, but characteristics of phase trajectories in that set *might* be significantly different than those of any nearby computable phase trajectories which could be effectively simulated. These characteristics include period, at the least, and—due to SIC—particulars of the areas which individual trajectories may visit even *within* a neighbourhood out of which the Shadowing Theorem may tell us they may not stray over a given time interval.

This is just the implication of the second observation, the observation that if a phase trajectory ever includes a computable point at a computable time, then it passes through computable points at every computable time. In a manner similar to the first observation and its followup, the second observation strictly partitions phase trajectories into

a computable set and a noncomputable set. And again, there are uncountably infinitely many phase trajectories in the second set but only countably many in the first. The interesting contribution of chaos is that because of sensitive dependence on initial conditions and topological transitivity, we are not guaranteed there exists a nearby non-diverging (as distinct from effectively converging, which we obviously are not guaranteed)[95] computable phase trajectory for any of the noncomputable phase trajectories. Thus, while there might exist particular pairs of phase trajectories which do not diverge from each other, there could be no general mapping of noncomputable phase trajectories to non-diverging computable ones.

Next we explore briefly the implications of these observations in the context of recursion theory in general as well as in the context of analysing the behaviour of the human brain as a dynamical system.

15.4 Computability and Behaviour—So What?

As they relate to the theory of computability itself, these observations suggest that the computability of a real-valued function is not sufficient to guarantee that algorithmic simulation of an analogue dynamical system governed by such a function could capture all of its interesting potential behaviour. The discrete representation required for algorithmic simulation essentially restricts our ability to simulate effectively all possible unique time evolutions of a chaotic analogue system which could take on an uncountably infinite number of states (not all of which are necessarily distinguishable, of course). Their sensitivity establishes for chaotic analogue systems a set of phase trajectories for which all possible corresponding computable phase trajectories not only fail effective convergence but are actually divergent. Perhaps a more rigid recursion theoretic taxonomy is required to distinguish between computability for functions and satisfactory simulation of real dynamical systems governed by such functions.

---

[95] This is a crucial but subtle point which perhaps should be emphasised. In any analogue system, we are already (trivially) denied the possibility of computable trajectories which effectively converge to noncomputable trajectories, but in chaotic analogue systems, we are denied the possibility even of computable trajectories which fail to diverge. And, of course, failure to diverge is a far weaker property than effective convergence!

Our conclusions are in the spirit of the comment of Vergis, et al (1986) who, after proving for a limited case of well behaved linear analogue computers a restricted form of the so-called Strong Church's Thesis—the thesis that any finite analogue computer can be simulated by a digital computer with resources bounded by a polynomial function of those used by the analogue computer—suggest that "any interesting analog computer should rely on some strongly nonlinear behavior". (p. 93) In addition, I believe these conclusions are not incompatible with the so-called Shadowing Theorem or the existence of chain recurrent sets. In particular, the existence of chain recurrent sets in no way implies the existence of computable tests for set membership, and *it does not imply that all members of such sets are themselves computable numbers.* Indeed, it has already been shown (Pour-El and Richards 1981, 1982) that the "computable" three dimensional wave equation of classical physics can transform computable inputs into noncomputable solution values (although Vergis and colleagues do note that this behaviour may be suppressed by imposing a continuity condition on the second derivative.) The ramifications of such nuances and the significance of continuity for recursion theoretic questions about chaotic systems in general looks to be an area ripe for more exploration.

In terms of application, the present conclusions suggest potential difficulties for attempts to analyse the capabilities of chaotic analogue subsystems of the human brain by observing the behaviour of downsized algorithmic simulations. The reasons why biological neural networks should be analysed as analogue rather than digital systems are complex, but for the present purposes it will suffice to note that while it is true that a single neuron either fires or does not (thus making it a simple on/off indicator), it is the spiking *frequency* which appears to be a primary carrier of information.[96] (A nice overview is included in Gustafsson, et al 1992. For applications to artificial neural networks, see the same or Gluck and Rumelhart 1990, Duchateau and Lansner 1991, Kong and Kosko 1991, Kosko 1992.) This frequency, unsurprisingly, is a continuous value. Other continuous parameters related to the behaviour of individual neurons are found in the mechanisms behind spike frequency adaptation and threshold evolution (see Shepherd 1990; also Nadel, et al 1989 for various

[96] The issue of simulating biological neural networks as analogue systems and the recursion theoretic ramifications were first discussed in technical detail in Mulhauser (1994a).

relevant data), as well as in the low level responses of ion gates themselves (Stühmer 1991, Neher and Sakmann 1992), if the "stochastic" behaviour of these latter are understood as manifestations of underlying continuous deterministic chaotic dynamics.

Despite years of comparative neglect in neural network research, the work of Freeman and others pursuing related research suggests that the special properties of chaotic dynamics have an important part to play in the way biological networks function. In particular, Freeman has postulated an explanation of olfactory pattern recognition in terms of the coupled nonlinear differential equations characteristic of chaos theory. Freeman's apparent discovery of strange attractors in olfactory cortex EEGs is often understood to indicate chaotic dynamics in low level communications between neurons.

The interesting question is whether a digital simulation of chaotic cortical areas could function as efficiently as its biological counterpart, given that an uncountably infinite set of possible phase trajectories available to the latter would be inaccessible to it. In the case of the olfactory cortex, it is interesting to wonder whether its pattern recognition capabilities are in any way dependent upon or aided by evolution through a specifically continuous domain enabled by the essentially analogue neural architecture. Similarly, we might wonder at the potential of chaotic analogue networks for solving NP-complete problems (Garey and Johnson 1979) in polynomial time, as Hopfield and Tank (1985) have explored—although their results are only good sub-optimal approximations. Promising results in this area have also emerged from research on magnetic alloy systems called spin glasses (Barahona 1982, Johnson 1983) and the related simulated annealing heuristic (Kirkpatrick, et al 1983; Aarts and Korst 1989). Vergis and colleagues (1986, p. 111) are able to predict only that spin glasses require an exponential settling time in the *worst case* over all inputs; whether or not these systems or chaotic analogue networks will yet emerge as contenders for polynomial solution of NP-hard problems remains to be seen. The bearing of our own explorations on such questions seems to be an open problem.

Questioning the algorithmic nature of processing in the human brain has become taboo in technical circles, largely in reaction to poorly founded arguments from philosophy of mind such as those put forward by Penrose in his controversial 1989 book. But setting aside the almost

religious faith of some AI researchers in the computability of all cognitive functions as well as some philosophers' similar faith in the noncomputability of at least some aspects of intelligent systems, it appears to be a question as yet unresolved whether genuine technical problems of a recursion theoretic nature may arise for analyses of the powerful processing capabilities of biological brains.

In the next three chapters we address a line of thought which is relevant to questions about the very existence of chaos in the real world and thus to whatever conclusions we might want to draw about conscious experience supervening on putatively chaotic neural substrates.

16

Chaos and Infinite Intricacy

We have discussed a number of points about the appearance of chaos when neural networks are analysed in the dynamical systems framework, and we have seen some of the ways in which chaotic characteristics might bear on the instantiation of self models. But these observations are vulnerable to attack in the spirit of a number of philosophical points Peter Smith has made recently in his Easter Term 1993 Cambridge lecture series on chaos. Smith's overall project seems to be to tranquillise some of the apparent philosophical hysteria associated with chaos theory as a "revolution" in scientific methodology and explanation. Like quantum mechanics, chaos science seems to lead otherwise sensible people to say some altogether unfounded things which are at best poor extrapolations from facts and at worse downright manipulations of them. I think something like Smith's overall ~~overall~~ project is desperately needed, and hopefully his points will have a purifying effect on philosophical discourse about chaos. But a couple of his comments, in particular those on mathematical models with infinite intricacy, on predictability, and on complexity (together with the relationship between complexity and representation) might appear to reduce to insignificance a number of the points we have made up to now. We must now follow Smith on a somewhat lengthy digression into philosophy of science in order to evaluate his analysis and the impact of points he raises in these areas on our own discussion of the significance of chaos theory for philosophy of mind.[97]

---

[97] Some of what follows might smack of an extended personal debate with Smith, but given that he has established himself as one of the leading authors in a very small pool of philosophers writing on such topics, it is only natural that the directions he has taken may bear significantly on our own project.

16.1 Chaos—Is it Really You?

Smith simplifies his discussion of chaotic systems as models of reality with a clever but inadequate line of argument. We might sum it up simply with the maxim that where there is no such thing as infinite physical detail, there is no such thing as chaos.[98] More precisely, he observes that the kinds of macroscopic physical systems to which chaotic models are typically applied are the kinds of systems which cannot truly exhibit infinite intricacy. Thus, he suggests, the characteristic feature of chaotic models (i.e., their infinite intricacy) cannot truly represent features of the real physical systems being modelled. (Smith 1993, pp. 8-9)

Smith wonders: can such a mathematical model with infinite intricacy really be a good model of a physical world in which there apparently is no infinite intricacy? How can an *infinite* amount of excess detail which is not borne out in the real world make for a good model? He points out (1993, pp. 10-11) that we apply models rather than equations to the real world and that what we are really after are models which are isomorphic to the real world. The equations, he says, are really just a way of specifying the model. If this is true, it seems to follow that whatever simplicity we might discover in terms of the *equations* of a chaotic model is of only minor importance in the face of such a disparity between the detail in the model and the detail in the world. And if there really isn't any chaos in the real world, then it seems that everything we have explored so far concerning chaotic dynamics in real neural networks may be completely irrelevant.

16.1.1 Intricacy, Real and Abstract

There is only a little to say about the presence of infinite intricacy in chaotic models. But Smith has two different concerns in mind about the absence of such infinite intricacy in the real world. The first, which he mentions only briefly, relates to quantum indeterminacy. The second derives from the fact that the macroscopic objects of most systems to which we would like to apply chaotic models are essentially abstractions such as centre of mass or velocity of a fluid. Presently we will take a quick

---

[98] Smith has objected in private communication that he is not denying the existence of chaos but merely challenging us with the question, "how can we use chaotic models when there is no true infinite intricacy in the world?"

look at the first of Smith's concerns and a more detailed look at the second. But first, as a brief aside, it is useful to make some observations about infinite intricacy in models themselves and the fractal nature of the strange attractors common to chaos theory.

In this area, Smith places far too much emphasis on fractals. The three properties to which we've appealed to characterise chaotic systems (and which define their own kind of infinite intricacy) often result in sets of points with a fractal character which are invariant for the *equations* of the mathematical model. But these fractal sets, or strange attractors, are mathematical *abstractions* in phase space. It verges on incoherence even to speak of fractals (or, as Smith has suggested privately, to speak of dense sets of periodic points, etc.) as anything *but* mathematical abstractions. Irrespective of loose wording from Benoit Mandelbrot or anyone else (see Mandelbrot quoted by Smith, p. 18), we should not expect fractals to have any real existence whatsoever. No one has ever seen a fractal, and no one ever will. Infinitely intricate mathematical abstractions do not exist on colourful computer screens, along coastlines, in EEG activity, or anywhere else. Fractals cannot even be observed (only *inferred*) in mathematical models in a finite amount of time. If anyone should ever approach you in a dark side road and offer to sell you a fractal, don't buy it. What they are selling is, at best, an approximation to a fractal which, *if* you could allow it to develop for an infinite span of time, *would* be a fractal. (Note that Smith follows this same line of reasoning with respect to the physical world, yet he makes very little of the fact that mathematicians don't see fractals either.)

But then, no one has ever seen a perfect ellipse either; they don't exist in planetary orbits, in geometry texts, or anywhere else. No one has ever seen the number e or the number π. We cannot conclude from the lack of perfect ellipses that general relativity is wrong, and from the lack of e and π running about, we cannot conclude that $e^{i\pi} - 1 = 0$ is wrong, nor can we conclude that somehow the circumference of a circle (if real circles only existed!) isn't really *exactly* the circle's diameter multiplied by π. And we certainly can't conclude from the absence of fractals in either the mathematical or physical world that coupled nonlinear differential equations don't *perfectly* describe the underlying mechanisms determining the behaviour of some types of physical systems. We will return to this kind of point more carefully again, but for now suffice to say

that from here on, we will concern ourselves primarily with whether the equations of a mathematical model can display the three characteristics which define a chaotic system and not with whether the physical world can display the impossible.

16.1.2 Intricacy at the Quantum Level

I said before that Smith was concerned with infinite intricacy both at the quantum level and at the level of macroscopic abstractions such as centres of mass and so on. We turn now to the first of these concerns. It is not unusual for philosophers to throw about the indeterminacy of state vector reduction as the catch-all fuzzy background for the whole world.[99] Almost anything might be precise, but as soon as we get to the quantum level, we are lead to believe, everything gets smeared out and imprecise and uncertain and fuzzy. For better or worse, this is only partly true. It is true that the results of our *measurements* of quantum systems are probabilistic in nature and that there is a fixed limit, quantified by Planck's reduced constant, to the precision with which we can know the values of two orthogonal observables. But while these features are undeniably part of the quantum landscape, it is often overlooked that unitary evolution of quantum systems in accordance with the Schrödinger equation is *entirely* deterministic and precise.

The Schrödinger equation is a continuous real-valued wavefunction, and while parameters of a system such as energy or charge may be a quantum, or discrete, value, this does not preclude there being an infinite class of, for instance, possible *positions* for a particle as described by its wavefunction. Just because the outcome of a transition from unitary evolution through state vector reduction is described probabilistically does not in any way mean that there is not a certain kind of infinite amount of possible detail at the quantum level courtesy of unitary evolution. Of course I do not mean to say we can build fractals out of particles described by wavefunctions, but quantum theory just does not pose any problem for the kind of infinite detail characteristic of the defining properties of chaotic systems, such as sensitive dependence on initial conditions. While we cannot say that such and such an arrangement of physical particles is a fractal, quantum mechanics in no way prevents us from

---

[99] This shouldn't be taken to imply that Smith is one of these philosophers! But since he has introduced the topic, we will grind the axe for good measure anyway.

saying that given any possible location in spacetime for a particle there is an infinite class of other points in the neighbourhood of that location such that *if* the particle were at one of those locations instead, it *would* follow a phase trajectory which diverges from the one it follows from the given location. (Note that this amounts to a sensitive dependence on initial conditions relative to a manifold of particular position where the values of orthogonal observables such as momentum could vary. This is a problem with trying to apply phase space descriptions to the quantum world, where we should properly be speaking terms of Hilbert space. The inaccuracy of the phase space description does not compromise the general point we are making, however.) It would appear that this will remain true unless quantum mechanics should embrace a quantum picture of spacetime itself.

16.1.3 Intricacy at the Classical Level

Let's turn now to the stickier question of infinite detail in macroscopic abstractions such as centre of mass or fluid velocity. Smith uses an example from celestial mechanics in which we are interested in the motion of a planetary object in terms of its centre of mass. He correctly observes that the actual centre of mass is a vague point because at any given moment it is indeterminate what particles should be counted as part of the planetary body. This brand of indeterminacy might be very small, and we might rightly count it as irrelevant to the kinds of questions we would want to ask about the planet's motion, but as long as there is any vagueness at all, Smith is right in concluding that it cannot exhibit behaviour which is infinitely intricate. Essentially, if we want to paint an infinitely detailed picture we need an infinitesimal paintbrush, and unless we can shrink the centre of mass to a precisely defined point our paintbrush will always be too large.

The problem does not go away if we treat this kind of vagueness as a problem of knowing the location of a point which really is there but which we cannot locate. It is tempting just to say, for instance, well, this planet's centre of mass is *here*, plus or minus a possible error of 5cm in all directions. In that case we would be interested in tracing the time evolution of the centre point of our area of uncertainty. Without assuming any kind of hidden variables interpretation of quantum mechanics, we could draw a rough analogy with the case for quantum

indeterminacy. To use again the example property of sensitive dependence on initial conditions, we might say that the centre of mass of a planetary body was somewhere within this cloud of indeterminacy, and wherever it might turn up for this particular observation, there is some other place arbitrarily close to it where it might have turned up instead and which would lead the planet along a diverging phase trajectory.

But this is inadequate for the reason that there just *isn't* in most cases a well-defined centre of mass. It's not a problem of ignorance, of there actually being a centre of mass which is just difficult to find. It is a problem with the centre of mass abstraction itself: just as Smith observes, at any particular moment it is just *indeterminate* which particles we should or shouldn't include in calculating a body's centre of mass.

But how much can we make of this kind of vagueness? Smith suggests that since we can't even decide where a centre of mass should be, perhaps we have no business trying to explain such a point's behaviour through a mathematical model which pins it down not just very precisely but can pin it down with an infinitely intricate relationship between its initial location and its location at some future moment of time. Isn't it a bit like trying to track the evolution of a nation's GDP without being clear on, for instance, what kinds of output should be included or whether we should include industries from the nation's territories and protectorates? I think there is something about applying infinitely intricate models to vague abstractions which so far has escaped scrutiny.

16.2 Intricate Models Wanted—Apply Within

Specifically, infinitely intricate models might still apply to any appropriate—perhaps even arbitrary—sharpening (pace Dummett and others) of vague abstractions such as centre of mass. In other words, we might say that *if* there were a well-defined centre of mass precisely at such and such a point, *then* that point would evolve through time to this other location in phase space. This is akin to saying that *if* a hill of sand must contain at least 4 million of grains of sand entirely within and not touching the border of a 15 cm radius, *then* after such and such a wind has blown on this hill of sand for a particular precise length of time the hill will have blown away. In practice, our definition of "hill of sand" is even more vague than our definition of centre of mass, but that doesn't mean it

would be useless to have a mathematical model which let us know how long it took for a hill of a given size to disappear under a given wind condition.

For macroscopic objects on the scale in which we're interested, the vagueness of abstractions such as centre of mass or hill of sand lies entirely in the abstraction and *not* in the real world. If we want to calculate something about the behaviour of such an abstraction with dynamical models, we must give the model something *real* and precise to chew on. When the model has digested the real input and given us back a description of behaviour, we cannot say the output is useless just because it was gotten from a sharpened real input which doesn't match our normal vague use of the abstraction; we must simply apply our vague abstraction *again* and recognise that the centre of mass or whatever is, as we will use it, more blurry than the answer we've got. The model is like the monster who chews up nails and spits out tacks, even if all we're wanting to know about is bumpy blobs of unfinished steel. Essentially we consider a vague quantity such as the location of a centre of mass, sharpen it up and feed it to the mathematical model, and then keep in mind that the output itself—like the real world—remains too sharp for our normal use of the abstraction and must be blurred to bring it in line with what we might actually observe. (Note that this is exactly analogous to our earlier conclusion drawn from the thermodynamics of black holes that we are applying continuous models to physical systems with only a finite number of distinguishable access states. The point holds *generally* and is not dependent in any way upon considerations peculiar to chaos.)

16.3 Ontological Significance of Models

The reason we must resort to such an inelegant interpretation of mathematical models of things like centres of mass is that we really shouldn't expect there to be *any* natural laws governing the behaviour of quasi-classical centres of mass. We *should* expect there to be laws governing gravity and the strong and electroweak forces which influence all the particles that go into making a body with a centre of mass. But looking for a natural law governing an abstraction like centre of mass for a macroscopic object is like looking for a natural law governing the behaviour of flying mallets. We might *derive* approximate models of

flying mallet behaviour by appeal to fluid dynamics (which of course is also concerned with abstractions), but mallets are as ill-defined as centres of mass. To apply a precise model of mallet flight to the real world, we would have to give the model a well-defined mallet, the likes of which we would never observe in the real world, and interpret what the model had to say about our abstract mallet according to the kinds of vague flying mallets we might actually observe. That this is necessary isn't a fact which should be blamed on the precision of our model, it should be blamed on the messiness of our flying mallet abstraction.

16.3.1 Limited Precision and Realism

Now someone might object along the lines of Smith's approach that even if we actually *could* use an infinitely intricate model to describe the time evolution of vague abstractions just by appropriately sharpening inputs and blurring outputs, there would remain a whole class of models *without* infinite intricacy which would offer at least as good or perhaps a better isomorphism with what we observe in the real world and abstract from it. That is, why should we bother with an infinitely detailed model when we're just going to throw away all that detail when we actually apply the model? Our answer here will be something that Smith has already dismissed as of no great importance, but I believe his dismissal was too quick.

We have noted already Smith's observation that models are what really get applied to the world, that equations are merely a way of specifying the model. But I suggest that what we are really after are the equations which somehow really *do* govern the motion of dynamical systems. If there are no equations really governing flying mallets as abstractions, we want the equations governing *sharpened* mallets as well-defined constructions of particles for which there *are* governing equations. Some of this begins to sound like a comment on realism. Smith mentions as an aside in several places that his view might be taken to imply some brand of anti-realism, but he does not take up the issue fully. In order to get a handle on whether apparently gratuitous infinite detail in chaotic models makes them inferior models of the behaviour of vague abstractions lacking infinite detail, we must briefly engage questions of realism edge on.

I propose that there *are* equations which precisely describe the motion of at least some kinds of bodies in the world. Whether or not the equations of quantum mechanics as it presently stands are a representative sample of such, I believe that there is *some* body of equations more or less like quantum mechanics which would precisely describe the dynamics of some kinds of bodies (but not of ill-defined flying mallets or vague centres of mass). This is little more than saying I believe the world operates according to particular laws of Nature. So far this is not a particularly stunning comment on the realism vs. anti-realism issue.

The implication of such a view is that there could be equations governing the behaviour of at least some bodies which are in fact sets of coupled nonlinear differential equations which can exhibit chaotic behaviour. This is independent of whether we have ever actually *observed* real systems governed by the equations exhibiting something like sensitive dependence on initial conditions. (Indeed, it would be impossible to *observe* sensitive dependence on initial conditions either in the real world or in mathematical models since we would need to try experiments over an infinite number of phase space points.) This is not as strong a realist position as it might at first sound; there is nothing startling about saying that even if all humans were born without any way of directly observing light Maxwell's equations might still be true. Maxwell's equations might still explain many things we *could* observe. In the same vein, although we cannot observe something like sensitive dependence on initial conditions, chaotic equations of motion might explain some of the behaviour we do observe through vague abstractions such as centre of mass.

This still does not answer the objection, however, that some other non-chaotic equations might just as accurately for our purposes describe the phenomena we are able to observe. But surely we are concerned not just with whether two candidate models, perhaps with appropriate sharpening or blurring or both, can each describe what we see. We are also concerned with the simplicity of the explanations the two models offer. Suppose for instance that someone offered a theory of planetary motion in terms of superpositions of various geometric constructions which, while highly complex, yielded a picture of planetary motion quite close to the kinds of motion we could actually observe. Suppose that someone else offered a simpler theory of planetary motion which didn't require such

complex constructions but which yielded a picture of planetary behaviour similarly in keeping with what we could actually observe. When physicists finally finally gave up trying to make complex epicycles work, it was both because they couldn't make the discrepancies with observed planetary motion disappear *and* because of the vastly greater simplicity, or parsimony, of the alternative. I would argue that even if there were *no* conflicts with observed planetary motion, perhaps because astronomical instruments weren't sufficiently capable, or perhaps even because there was some limit *in principle* to what they could measure, we should *still* count epicycles out in favour of simpler alternatives, whether Newton's or Einstein's. Likewise, even if there are other models which describe, for instance, the time evolution of a vague abstraction such as centre of mass, without involving infinite intricacy and the attendant sharpening and blurring, if these two models are in practice indistinguishable, we should opt for the one with simpler equations.

The comment on realism then is simply this: that one set of equations or natural laws is *correct* as applied to appropriately sharpened abstractions[100] and that we are then justified in applying standards such as parsimony in choosing one model over another experimentally indistinguishable model. And as far as I can see, the simplicity and elegance of the equations of chaotic models is unlikely to be surpassed by another model which gives up the power of deterministic coupled nonlinear differential equations for the sake of avoiding infinite detail. If anyone should produce a model of a physical system which *does* do this job, then we would be well advised to opt for it over a chaotic alternative. Shortly we will discuss the suitability of (often computationally simpler) stochastic models for applications where we are not interested so much in what actual equations govern physical systems but only want a model which displays behaviour similar to that of the real physical system. But first we visit again the recurring theme of description levels.

16.3.2 Limited Precision and Levels of Description

All this so far is well and good, but there still is one more twist in the story of intricacy in models of vague abstractions such as centre of mass. Recall our earlier discussions about tracing time evolution of

---

[100] And with a cosmologist's bent, I mean here *noise free* abstractions; we return to this question near the end of this chapter and again in Chapter 18.

neural networks at different levels of description. There we noted that dynamical behaviour is underdetermined at the ω and ψ levels, despite quasi-classical deterministic λ dynamics. The rationale behind this point was that several points distinct in λ space might be included in the same point in ω or ψ space and that these points might lie on phase trajectories which diverge not only in λ space but also in ω or ψ space. There is a problem analogous to applying mathematical models of limited intricacy to abstractions like centre of mass. The problem can be understood either as a difficulty with using deterministic mathematical constructions to model physical phenomena described at a level that might be underdetermined or as a difficulty with unduly reducing the dimensionality of the model system to below that of the real system being modelled.

Returning to the example of centre of mass as it relates to the first way of understanding the problem, we can note that an infinite number of possible distinct physical arrangements (not all of which will be physically distinguishable) yield the same centre of mass for a macroscopic object of a given approximate mass, shape, and density. Thus, if it should happen that any characteristics of the actual position/momentum distributions of the particles in the macroscopic body (beyond the rough information given by the centre of mass) are ever relevant to that body's dynamics as understood through the centre of mass abstraction, then any model which doesn't track these characteristics will be inherently subject to error. Indeed, if the actual equations governing the dynamics of the particles in the macroscopic body (whether we know these equations or not) are chaotic, these errors may ultimately become quite large. Thus even if we cannot in practice observe fine details of the structure of a macroscopic body, a mathematical model which lacks the (in practice, unobservable) intricacy of the real physical system may not describe the behaviour of what *can* be observed precisely enough to be useful. Without exploiting the fine details of a system at a level of description *below* the level of what we are actually wanting to describe, a model may not be able to account for what would otherwise appear to be indeterminacy at the higher level of description. This is just another version of the earlier observation that there is unlikely to be a deterministic description of ψ level dynamics which works solely with ψ level information.

Another aspect of this kind of problem arises from the (often necessary) reduction in dimensionality as we move from the real world system to its model. When we model a system involving an abstraction such as wind velocity or air pressure, we are in effect reducing the dimensionality of the model by counting the abstraction as one single variable rather than as a composite of variables each describing aspects of individual gas particles. Rather than causing a problem with our long term qualitative description of the behaviour of the abstraction (the air pressure or whatever) as in the above, this way of reducing dimensionality *does* mean that we can no longer legitimately maintain the uniqueness of phase trajectories of bodies moving in the gas. Essentially, we are *projecting* higher dimensional dynamics onto a manifold of lower dimension. This may be entirely appropriate when we really don't care about the system's dynamics in all those higher dimensions (such as all the motions of the individual particles of gas as opposed to the simple abstraction of wind velocity or some such). But this also means that phase trajectories unique in the higher dimensional space may actually *cross* in the lower dimensional projection. Thus, even dynamics tracked at an intricately detailed low level may still be nondeterministic at that level. The system as plotted in the phase space of the dimensionality in which we are interested might pass through the same point over and over again while never passing through the same point in a higher dimension. Thus, in the lower dimension, the system may eventually cease passing through the same point and move into some other regimen, while remaining entirely deterministic.

The usual way of coping with this observation is to say that the system being modelled is subject to environmental noise. When the dynamical influence of the higher dimensions is relatively unsystematic, it is easy enough just to add it in as noise, and the behaviour of the model will be qualitatively similar to that of the real system. But this is essentially a clever *fix* for overcoming the basic problem with the model: that it is does not pay enough attention to the intricate detail of the real world. This is not to say it isn't a wise thing to do! Tracking the intricate detail of a system in millions of dimensions is practically impossible and a waste of resources if we can get the same results by injecting a little noise. But what we are doing is just making up for the vagueness of our abstraction (wind velocity, air pressure, or whatever) by adding back in an

influence that we have established empirically to be rather unstructured and statistically uniform. (We will return to this point in more detail in the chapter on complexity.)

16.3.3 Ontological Significance in Practice and Theory

This brings us to the final point about modelling vague abstractions or high level features with intricately detailed chaotic models. In those cases where the influence of very low level dynamics on higher level abstractions is unstructured and uniform, it would seem we are perfectly justified in using what may be computationally simpler stochastic models for certain applications. For *practical purposes*, stochastic models may be computationally less intensive than highly detailed chaotic counterparts, and they may display nearly identical qualitative behaviour at the levels of description in which we are most interested. But in terms of understanding how the physical systems *really* work, as in the above discussion of realism versus anti-realism, the chaotic models are preferable because they are deterministic. Indeed, there is a strong case for the idea that all physical behaviour which can be interpreted as stochastic, perhaps with the exception of probabilistic state vector reduction, is really, at a lower level of description, deterministic chaos. A full exploration of such an ontological position is unfortunately beyond our scope for now.

But before we finish with this topic, it is useful to notice that many of the chaotic models we actually use are not *really* models of the kinds of systems we actually observe in the world. As Smith notes in several places, real observable systems are always subject to environmental noise. But the equations for chaotic models we actually use often have no terms for this noise. Our equations specify models of what we might observe if there were no noise. That there are no real systems where noise is entirely absent—save the entire cosmos—is not a criticism of these models, because a correct model may perfectly well describe "what makes a system go" independently of *external* noise. A model of what makes a car go may describe perfectly well an engine and a transmission and a drive train and so on, but it may ignore entirely the fact that no real cars ever go precisely as the model predicts, because there is always air turbulence and road irregularity and so on. But these latter aren't what makes the car go! And it is no criticism of a model which ignores air turbulence and road irregularities that it doesn't perfectly match the behaviour of observed

cars, even if some other very complex and perhaps stochastic model *could* account for the air turbulence and road irregularities and their influence on observed reality. Moreover, few have ever called into question the explanatory power of Newtonian mechanics or General Relativity or quantum theory just because none of their equations include terms for environmental noise!

Likewise, if we are interested in ontological questions about what makes things go, instead of just good engineering sense when it comes to making good predictions about the world, Smith's objections about a lack of infinitely intricate behaviour are irrelevant. To put it succinctly, Smith suggests that if an infinitely detailed model yields, for instance, fractal structures in phase space, but there are no fractal structures in nature, then the model cannot be an accurate description of reality. But a model's inability to account for the influence of countless tiny perturbations *external to its domain of explanation* just isn't a criticism of that model's description of reality. And it is no more a criticism of chaotic sets of equations that they have fractal invariant sets when no such fractals are ever observed than it is a criticism of an equation yielding the RPM of an engine at a certain throttle setting that no car running down the road ever has it's engine running steadily at that RPM for the given throttle setting.

16.4 Models Through the Intricate Haze

We have addressed, then, Smith's concerns that chaotic models are not suited to the kinds of systems to which they are typically applied. We have noted the inappropriateness of paying too much attention to the infinite intricacy of fractal attractors rather than the kind of intricate intricacy we see in the three defining characteristics of chaotic systems, and we have discussed the sense in which for understanding how systems "really" work, chaotic models may well be the best candidates. Smith had challenged that perhaps where there was no infinite intricacy, there was no chaos. But hopefully the situation is somewhat clearer now and we can see that chaos is alive and well and living in the real physical world. Now we must move on to a discussion of Smith's comments on predictability as it relates to reality and chaotic models and see if they have any more significance for our purposes than his points on infinite intricacy.

It may not be necessary for a model of a chaotic system such as a particular neural network to be completely predictable in order to exhibit the kind of useful behaviour of interest to the cognitive scientist. Indeed, it may be only necessary that it display qualitatively similar behaviour. After all, there are no perfect simulations of *particular* human brains, but there are plenty of human brains out there with behaviour which is qualitatively similar to each other and which have properties of use to the cognitive scientist. But the question remains whether chaotic systems have any particular properties relative to predictability which may be relevant to simulations of systems like complex neural networks. In what follows we will discuss the ways in which chaotic systems *are* unique and demand special considerations for computer modelling. In particular, the low level chaotic behaviour of some neural networks might require very detailed low level simulation in order to extract behaviour which is sufficiently similar to biological reality, and it might suggest a way in which characteristics of subjective experience in the self model may be intimately related to the functioning of their instantiating substrate.

17

Chaos and Prediction

Broadly speaking, Smith's aim in discussing predictability is to pin down the sense in which he can say that chaotic systems are just as predictable as any other kind of system and not worthy of any special attention in this area. This is contrary to what seems to be the prevailing notion among philosophers that chaotic systems are altogether unpredictable and are unique among quasi-classical systems in so being. If Smith is correct, then chaotic features of the instantiating substrate of self models are unworthy of special attention in terms of predictability.

17.1 Quantifying Predictability

17.1.1 Epistemic Determinism

Smith begins by quantifying the notion of "predictability in principle", or "epistemic determinism" with the requirement [P1] (Smith 1993, p. 32):

[P1] $(\forall\delta)(\forall t)(\exists\varepsilon)$(the state after interval t can be fixed within δ by fixing the initial state within ε)

But [P1] is trivially satisfied by *all* physically deterministic systems, chaotic or otherwise: given any δ and any t whatever, $\varepsilon = 0$ satisfies [P1]. Since we are talking about deterministic chaos anyway, [P1] is entirely superfluous. Probably Smith had in mind something like [P1'][101]:

[P1'] $(\forall\delta > 0)(\forall t)(\exists\varepsilon > 0)$(the state after interval t can be fixed within δ by fixing the initial state within ε)

This formulation captures Smith's notion of "epistemic determinism" in a way that [P1] does not because determinism straightforwardly entails [P1] whereas it is at least less obvious whether it entails [P1']. Smith seems to intend [P1'] as just a comment on the relative stability of infinitesimally

---

101 Indeed, Smith has noted in later private communication that he was just simplifying the formulation by assuming $\delta > 0$ and $\varepsilon > 0$.

separated trajectories as quantified by the Lyapunov exponent λ (which is positive for typical trajectories of chaotic systems), where for typical trajectories divergence is proportional to $\varepsilon e^{\lambda t}$. In a moment we shall see that [P1'] is ambiguous and doesn't *quite* capture what we're after.

But first we can immediately observe one important point about applying [P1'] to real physical systems as opposed to mathematical models. The kind of "epistemic determinism" Smith has in mind with [P1'] (or his original [P1] perhaps) generally applies only to mathematical models and not to the real world. While there are plenty of cases where [P1'] holds for a mathematical model, in (almost?) none of these cases does it hold for the real physical system being modelled. The reason is very straightforward. Given quantum mechanical bounds on measurement, we can *never* fix a system's initial state finely enough to satisfy the $(\forall \delta > 0)$ quantification. Indeed, an infinitesimal δ here doesn't make sense beyond the scale of Planck's constant, but even when it is safely within that scale, given the exponential phase trajectory divergence common to chaotic systems, it certainly does not follow that the ε required to meet [P1'] will also be quantum mechanically plausible. Thus, even if we fix up [P1'] to take care of quantum mechanical uncertainty on the final state end, it remains false for physical systems whenever an impossibly precise ε is required. This is all entirely compatible with saying the mathematical model of the system meets [P1']. In order to formulate a criterion for predictability in principle as it relates to real physical systems, we might apply another fix to the [P1] series to make sure of quantum mechanical sense on both the initial and final ends of the matter. But I suggest that we keep with the same style for describing predictability and just keep in mind that real physical systems typically won't meet our requirements. There remains a minor refinement to be made to [P1'] which will change our characterisation of predictability to a form very similar to the Shadowing Theorem.

The wording of [P1'] or a derivative fixed to account for problems of quantum measurement does not allow us to distinguish different behaviour of systems in the vicinity of different points. In particular, our notion of "epistemic determinism" or "predictability in principle" should allow us to distinguish the sense in which predictability is limited in the vicinity of a what is called a singularity. (Note that in an earlier IPPE-distributed draft of a document including this chapter, I referred to predictability in the vicinity of *critical points*—a class of points of which

singularities are a special case—and after describing the example below suggested that they were common to chaotic systems. In this my discussion was entirely *incorrect*, and I am grateful to Peter Smith for clarifying my thinking on this point. While singularities certainly may exist in chaotic systems, they are *not* typical of such systems,[102] and Smith's original text was altogether correct in stating that something like both [P1'] and the following [P*] are met by typical chaotic systems.)

Consider a simple sensitively dependent but nonchaotic system consisting of a sphere of uniform density placed atop a cone, with a gravitational attraction at the open end of the cone and in line with its axis. If we follow for the moment a simple model of the system with no rolling resistance and a perfectly smooth sphere and cone, and we say the sphere will just come to a stop when it has rolled down the side of the cone and hit some perfectly smooth and energy absorbent plane at the base of the cone, it is easy enough to read off the sphere's final position just by noting the angle θ of the sphere's initial "north pole" in polar coordinates. In it's final position, the sphere will rest tangent both to the cone and to the plane, with its centre at the same θ. (With information about the initial ϕ as well, we could express the final position in terms of where that initial north pole has gone, but the extra detail is irrelevant for the present discussion.)

In this simple system, the initial condition corresponding to values of zero for θ and ϕ is called a singularity. If we operate with [P1'] as our description of what it means to be "predictable in principle", we can comment only on the overall behaviour of the system, completely missing out anything special about the singularity. It would be useful to be able to say something about the predictability of the system with respect to particular neighbourhoods. To do this, we must be able to speak about fixing errors around particular initial conditions rather than just about fixing errors in general. Smith already had this kind of description to hand in his earlier discussion of what is commonly called the Shadowing Theorem, but [P1'] doesn't capture all the subtlety of that discussion. [P*], which for our purposes can be thought of as a derivative of the Shadowing Theorem, does the trick:

---

[102] Lipschitz condition satisfaction and continuity of the equations describing a system—properties shared by the bulk of chaotic systems—guarantee no singularities.

[P*] $(\forall \delta > 0)(\forall t)(\forall x_0)(\exists \varepsilon > 0)(\forall y_0)$(if $|x_0 - y_0| < \varepsilon$, then after interval t, $|x_t - y_t| \leq \delta$)

Here we can see that [P*] fails specifically when x_0 = the singularity and δ corresponds to a difference in position smaller than half the circumference of the circle formed where the cone meets the plane (assuming the length of the path from the point of the cone to this circle is less than this amount). The reason is that, given x_0 = the singularity, any ε whatsoever includes y_0 points with θ angles separated by exactly π radians, and these initial conditions correspond to final states on opposite sides of the cone. Note that on a strict interpretation (presuming universal quantification over all possible initial states) [P1'] is false for this simple system and for any system with one or more singularities; such systems do not possess this sort of "epistemic determinism". We could save the idea by interpreting [P1'] more loosely but at the cost of ignoring singularities. [P*] remedies the shortcoming by incorporating the x_0 which reveals which particular neighbourhoods are home to trajectories which are "predictable in principle". Like a strictly interpreted [P1'], [P*] is false for any systems with singularities, but we at least have a way of specifying the set of x_0 for which it fails.

There are a number of things we might notice about predictability in the neighbourhood of singularities. Most importantly, we are not claiming that behaviour in the neighbourhood of critical points is nondeterministic (the systems can still meet Smith's original [P1]), nor are we claiming that predictive errors in the vicinity of such points are somehow unbounded. It's just that with respect to the space of possible ending states of systems like the sphere on the cone, the predictive error can be very large (in this case, the whole space). This is strictly compatible with more relatively stable (as quantified by the Lyapunov exponent) *typical* trajectories.[103]

17.1.2 Qualitative Predictability

Let's return for now to the question about magnitude of predictive error. For some systems, we are interested to know a system's "final state"

---

[103] In private discussions, Smith has suggested that when the set of critical points in a given system has measure zero, we are justified in throwing out that set because a randomly chosen point in the phase space of the system has probability zero of landing on a member of the set. But since the phase space volume of relevant *neighbourhoods* around such points may not have measure zero, I believe this is unjustified.

only in terms of what basin of attraction the system has entered. Large errors in predicting the exact future state of a system do not concern us, because we are curious to know just whether the trajectory is one of a set which will asymptotically approach an attractor in a particular area of phase space and exhibit behaviour qualitatively similar to that of other trajectories in the same area. Also, we might not be concerned with deciding whether a system actually will be within a certain distance of a given attractor; once it is in the basin of an attractor, then we know it will tend towards the attractor, however quickly or slowly, and this may be enough for our predictive needs. We might call a system for which predictions of this kind are possible "qualitatively predictable", in the sense of [Q], where B represents a basin of attraction:

[Q] $(\forall B)(\forall x \notin \text{boundary of } B)(\exists \varepsilon > 0)(\forall y)(\text{if } ((|x - y| < \varepsilon) \,\&\, (x \in B)) \text{ then } y \in B)$

Note that [Q] misleadingly appears to be a weaker description of predictability than [P*]: in any neighbourhoods where a system satisfies [P*], it looks like it must also satisfy [Q]. [P*] gives us, for any desired predictive accuracy over any time interval, a required initial measurement accuracy. [Q] seems to require, essentially, that for every point x not on a basin boundary there is always a measurement accuracy fine enough that we can be sure the neighbourhood within δ (from [P*]) of x doesn't overlap a basin boundary. Since the distance from any point not on a boundary to the nearest boundary is trivially positive, it seems [P*] can provide any ε we require to satisfy [Q]. Momentarily we will examine a system which illustrates how this line of reasoning is flawed.

It is clear, of course, that the converse relationship does not hold: a system's meeting [Q] in particular neighbourhoods does not imply its meeting [P*] in those same neighbourhoods; for a simple example where one is met but not the other, consider any system which has a singularity in an attractor basin (admittedly, this is odd, but it's certainly not impossible!). In such a system [P*] might fail in the vicinity of the critical points; yet trajectories in the same neighbourhood could never leave the basin and end up in a different one, and [Q] would be satisfied.

Now if a given initial condition, an x from [Q], lies far from the boundary of a basin of attraction, then the "final state" of the system can be predicted with *certainty* as long as our measurement error ε is smaller than the distance from x to the basin boundary. Even if x is close to an

ordinary (nonfractal) basin boundary (or, alternatively, if ε is larger), such that some of the points within a particular ε of x are actually in a different basin of attraction, this proportion of initial conditions with uncertain outcomes generally scales linearly with ε. In terms of extrapolating this property from mathematical models to systems in the real world, incidentally, it is worth noting that [Q] can no more be true of physical systems in general than can [P1'] and [P*]. (This is because we can choose an x close enough to a basin boundary that quantum uncertainty prevents us shrinking ε small enough that it doesn't overlap the boundary.) Where a model's qualitative predictability in the style of [Q] becomes most interesting is in the neighbourhood of a fractal basin boundary.

17.2 Qualitative Unpredictability with Epistemic Determinism

In this case, the proportion of uncertain initial conditions within ε of x scales *nonlinearly* with ε, in a way that is usually related to the Hausdorff-Besicovitch dimension of the basin boundary. In a correlate of Smith's observation that chaotic systems demand exponentially more accurate initial measurements as the desired time interval for prediction is increased, initial measurement accuracy requirements for prediction of final state in terms of attractor basins can also be highly nonlinear. Grebogi and colleagues (1987), for instance, describe a model of a kicked double rotor where decreasing the initial error ε by a factor of 10^{10} yields a decrease in the proportion of uncertain trajectories within ε of x of only a factor of 10. But even with such high costs in initial data, such a system remains qualitatively predictable in the sense of [Q].

17.2.1 Riddled Basins

Recently, however, John C. Sommerer and Edward Ott (1993) have described a chaotic model for which [Q] fails but for which, curiously, [P*] does not. In the Sommerer and Ott system, we are faced not just with a highly nonlinear relationship between ε and the proportion of uncertain points. The "final state" of their system in terms of attractor basins is uncertain for *every* $\varepsilon > 0$; moreover, this property holds not just on a limited fractal basin boundary (of the sort described by Smith pp. 32-33; more on this presently), but over the *entire* phase space volume of the system's two basins of attraction.

These basins of attraction are called *riddled*; an attractor basin B is riddled if it satisfies the following [R]:

[R] $(\forall x \in B)(\forall \varepsilon > 0)(\exists y)((|x - y| < \varepsilon) \& (y \notin B))$

In other words, a basin is riddled if, for any initial state x in the attractor's basin, the set of points within any arbitrarily small distance $\varepsilon > 0$ of x but not in the same basin as x has positive volume in phase space. The Sommerer and Ott model is the first example of a physical system (as opposed to just a mathematical mapping, as in Alexander, et al 1992) which possesses riddled attractors. The model describes a particle moving in the x-y plane with an acceleration given by the sum of three forces: the gradient of a symmetric scalar potential, linear friction opposite to the particle's direction of movement, and a periodic external force in the x-direction. They reduced the dimensionality of the system through a function that returns mappings of points on a Poincaré section at times in phase with the periodic external force.

The four-dimensional phase space of this simplified system possesses an invariant plane (along $y = v_y = 0$) with a strange attractor. Within this invariant subspace, one Lyapunov exponent of typical trajectories is negative while the other is positive, thus indicating chaotic dynamics on the attractor, and normal to the invariant subspace both Lyapunov exponents are negative for typical trajectories. This implies that a set of points not in the invariant subspace, with nonzero phase space volume, is attracted to the chaotic attractor, but, as the authors point out, this does not rule out a dense set of atypical orbits on the attractor having a positive normal Lyapunov exponent. This latter condition allows for initial conditions arbitrarily near the invariant subspace to be repelled from it. These points may eventually find themselves in one of two repelling regions of the scalar potential and go to positive or negative infinity in y and v_y, the y velocity component. Thus there is a second "attractor" at $\pm\infty$ along these two dimensions. Sommerer and Ott's work is the first known model of a physical system with the three conditions shown by Alexander and colleagues to imply riddled basins of attraction: an invariant manifold in phase space with a strange attractor, negative Lyapunov exponents normal to typical orbits in the attractor, and a positive volume of initial conditions in any region of phase space attracted to a different attractor.

Although this is the first model of a physical system with riddled attractor basins to be studied, and while such systems are of course only a subset of all chaotic physical systems, Sommerer and Ott indicate that the conditions for such systems are "by no means so restrictive that riddled basins can be considered unnatural". (Sommerer and Ott 1993, p. 140) Indeed, apart from the symmetry of the scalar potential, the equations of the system are unremarkable, and the system's riddled basins are a highly robust feature which do not disappear under a wide range of changes in control parameters. It seems unlikely that such systems will be rare in nature.

17.2.2 Riddled Basins and Qualitative Predictability

For our present discussion, the first immediately salient point about such systems is that riddled attractor basins are strictly incompatible with [Q]. This is straightforward: when [R] is true, there will be a y in a different basin of attraction to x for *every* $\varepsilon > 0$. Yet [Q] would require some $\varepsilon > 0$ which guaranteed that every y within ε of x was in the same basin of attraction. It appeared originally that satisfaction of [P*] implied satisfaction of [Q], and so we might infer that systems with riddled basins fail [P*] as well. Yet while such systems are not "qualitatively predictable", they do remain "epistemically deterministic" as we have defined the terms. The line of reasoning which suggested that [P*] implied [Q] was flawed in that it assumed there was some distance from the x_0 trajectory to the nearest basin boundary; but in the case of riddled basins there are "holes" leading to other attractors in *all* neighbourhoods, and [P*] cannot provide a δ to satisfy [Q] for a given x. There is no indication that Sommerer and Ott's system could somehow circumvent the Shadowing Theorem; thus we can only assume that along with a broad class of coupled nonlinear differential equations it does satisfy [P*].

Of course riddled basin systems are also still deterministic in the original [P1] sense: if the initial condition is known *precisely*, then the system's exact final state can, in principle, be calculated without error for any future time. But if there is any imprecision whatsoever in our knowledge of the initial condition, we retain the ability to predict the system's future state within δ for any chosen amount of time, but we lose the ability to know the system's ultimate behaviour in terms of attractor basins.

Thus, paradoxically, we can shadow the centre point of a neighbourhood to within an arbitrarily small distance, but we cannot comment on the destiny of other trajectories within that neighbourhood except up to the point in time to which we have already calculated. Essentially, we can keep trajectories within a certain area, but we cannot say where they are going! The best we could do would be to choose each of the points in the neighbourhood in turn and calculate their trajectories, one by one. Even if in principle we could calculate the destiny of any single point in phase space—a dubious proposition in itself, to which we will turn presently—we still could not infer anything about the long term destiny of points lying arbitrarily near it. Within the distance δ of where we might know a particular point x will map after a given amount of time, there is *qualitative* variance in trajectories in that a positive volume of points within ε of x will diverge from it sufficiently to go to a different attractor than x itself.

17.2.3 Riddles and Convolutions

Now, here's a curious fact about this kind of system which differentiates it from the example of the desktop toy suspended above three magnets which Smith discusses. (1993, pp. 32-33) In the pendulum example, there are convoluted boundaries between the basins of attraction such that a set of initial points near the boundary of attraction to one magnet is dense in sets of points near the boundaries of attraction to the other two and similarly for the sets of points attracted to each of the three magnets. Smith notes that "since the approach to the attractor is asymptotic, the "final" position of the pendulum must mean its position in the limit as time goes to infinity" (p. 33). That's fair enough, but (recalling that this is a dissipative system) if we observe any arbitrary nonperiodic trajectory there will be for it *some* time after which we can be sure the pendulum isn't going to leap over to some other magnet.[104] This is different, of course, to saying that there is some time which will do for all possible trajectories (which is like saying there is some largest integer).

---

104 This is unless we had a model in which the magnets are so powerful that asymptotic motion within all $e > 0$ of an attractor may still be "disturbed" and pulled toward another, but in the typical example there *is* some area in which the attractor basin is no longer convoluted (or, alternatively, "locally riddled"), and once a nonperiodic trajectory enters that area—which it must do eventually—we can be sure of where it is going.

But in the case of riddled attractors, we may observe any arbitrary trajectory for as long as we want, but unless that trajectory begins life on the invariant plane or in the repellent area of the scalar potential, (unlike the pendulum example) unless and until it enters the repellent area, we can *never* be sure where it is actually going. In the pendulum example, there is a region of phase space like a "safe house", such that once a trajectory gets there, we know it will definitely go to the attractor point at infinity. But in a riddled basin, we might never know if a trajectory is safe; it may still get sucked right back towards the other attractor. (Sommerer and Ott's work is the first model of a physical system which exhibits this feature.) This curious fact is important to the following section.

17.3 Riddled Basins and Computability Revisited

Recall that qualitative variance of trajectories within a small neighbourhood of a given point on a computable trajectory is exactly what we indicated earlier, on *a priori* grounds, should be possible for a class of analogue chaotic systems. We have observed already that due to problems of quantum measurement, our criterion for qualitative predictability is not met by actual physical systems which are chaotic. But we now have to hand an example of a mathematical model which also fails to meet the criterion. To find unpredictability in the physical world was one thing, but to find it in the mathematical world is another.[105] We have encountered a mathematical model governed by computable functions which displays an aspect of noncomputable behaviour.

It is specifically the continuous nature of the equations of motion Sommerer and Ott have exploited which yields exactly the kind of situation I earlier argued was possible, in which computable equations do not guarantee that all aspects of a system's behaviour are computable. We cannot actually know with certainty the destination of any point not on the invariant plane or in the repellent area of the scalar potential, but supposing we chose some distances which would do for "close enough" to the attractors, with only a denumerably infinite class of computable phase trajectories passing through points within a given ε of x_0 at a particular

105 On a related side note, see Chaitin (1994) for an example of a beautifully noncomputable number which emerges from basic algebra as well as for insightful comments on the "decline and fall" of reductionism in mathematics in the face of mathematical truths which are "true for no reason at all".

time, we could computably set about the task of calculating each of their ultimate destination in terms of (satisfactory proximity to an) attractor basin. But doing so still gives us absolutely no information about the uncountably infinite class of noncomputable phase trajectories passing through the same space in an analogue system. For all we could know, the "real" destination map of an analogue system with riddled attractors might spell out "© God, All Rights Reserved!" in noncomputable points in one attractor basin against a background of noncomputable points in another basin!

Indeed, as I noted above, we cannot even be sure of the *qualitative* accuracy of the destination maps Sommerer and Ott include in their article. Regardless of the host of problems associated with the roundoff error in calculating the trajectories of points Sommerer and Ott chose, while we can be certain that those points which definitely arrive in the repellent area of the scalar potential will be pushed to infinity, we cannot be sure that any other points even arbitrarily close to the invariant plane will ultimately be attracted to it.[106] Sommerer and Ott note this as the possibility of "arbitrarily long transients" but fail to note its implication for the qualitative accuracy of their destination maps. It is tempting to think that the Shadowing Theorem guarantees that calculating the computable trajectories would give all the information we needed and that no trajectories in the neighbourhood would behave significantly differently, but we have already noted that truth of [P*] with respect to a given neighbourhood does not imply that [Q] is also true in that neighbourhood.

Of course the complexity of the Sommerer and Ott system is bad enough without even considering questions about analogue systems: the dynamics of the system as run on a digital computer are highly complex as well! The specifics of the system's behaviour can be reproduced only by digital computers with identical architectures and numerical evaluation algorithms. The authors note that their calculated destination maps of points in the Poincaré sections they studied differed in fine detail (but not in general riddled character) when they used different computers with different round-off algorithms and precision. This just reinforces the

---

106 Sommerer and Ott colour points within a very short distance of the invariant plane ($|y| < 10^{-8}$, $|v_y| < 10^{-9}$, $y \bullet v < 0$) as if they ultimately went to the plane (p. 140), but unlike systems without riddled attractors, we still cannot be sure!

point that the details of behaviour for such a digitally simulated system would likely be vastly different than for any similar analogue system.

Although I did not have the [P*] and [Q] formulation to hand or even any knowledge of Sommerer and Ott's research when I originally outlined the earlier arguments about computability of chaotic analogue systems, the new system suggests a way of quantifying one route to noncomputable behaviour in terms of [P*] and [Q]. Simply, it would appear that any analogue chaotic system described by computable functions and which fails [Q] but still meets [P*] will display the kind of noncomputable behaviour we have discussed. (In complete fairness, however, the problems about noncomputable behaviour in the Sommerer and Ott system come largely for free as a result of the model's deep complexity and are not dependent on the train of logic I outlined in the original section on computability and chaotic analogue systems.) This is not to say there might not be other ways to get to similarly interesting behaviour, and the earlier arguments hint at what some of these might be; but this certainly appears to be one route to noncomputable system behaviour in the face of computable governing functions.

17.4 Prediction, After the Facts

We have explored Smith's analysis of predictability in chaotic systems and seen some of the ways in which they do remain different than nonchaotic deterministic systems in general. Although the systems with [R] style basins of attraction, for which [Q] fails while [P*] remains true for typical trajectories, are the most striking examples of properties unique to chaotic systems, we have also noted how [P*] may fail in the vicinity of some points and what impact this has on the systems' predictability. The unique properties of chaotic systems are such that they cannot be quickly dismissed as irrelevant to cognitive modelling just because chaotic systems are in general "predictable in principle" like other deterministic systems. These unique properties may have functionally relevant rôles in real intelligent systems which will not be served by any but the most detailed simulations of extremely fine physical details of neural networks present at or below the level of individual neurons.

Next we consider the problem of complexity and relating chaos to simple noise.

18

Complexity Simplified

We turn now to Smith's last comment on chaotic models which is significant for our own project. The issue this time is in what sense, if any, chaotic behaviour can be categorised as random. A central question is whether the apparently chaotic behaviour we observe in the real world is qualitatively the same as the chaotic behaviour we observe in mathematical models. In the mathematical models, Smith suggests, detailed chaotic behaviour might be viewed as the result of the equations' nonlinearity magnifying the fine details of the expansions of real numbers used to specify a system's initial conditions. In the real world, Smith observes, chaotic behaviour might be viewed as at least partially the result of the continual influence of environmental noise. We noted above, in the discussion about describing the behaviour of physical systems with mathematical models of lower dimensions, that injecting noise is a common and often indispensable tool for bringing the model's behaviour in line with what is observed in the physical world. But if it is more than a tool, if apparent chaos in the real world is qualitatively as random as noise (whether it really originates just from noise, just from chaos, or from some combination of the two), then almost all philosophical questions we might ask about the importance of chaos *qua* chaos dwindle to irrelevance. Now we must examine more closely the rationale behind equating the "randomness" of low level chaos with that of noise. The first issue is how we may quantify complexity in order to compare chaos and noise, and it is to that which we next turn our attention.

18.1 Defining Complexity

18.1. KCS and Shifty Business

Smith begins his exploration of the issue with an appeal to the Kolmogorov/Chaitin/Solomonov, or KCS, definition of algorithmic

complexity. Algorithmic complexity, also known as algorithmic information content, algorithmic randomness, or algorithmic entropy, can be usefully applied to physical entities, rather than just bit strings as we will use it here. It is the length, in bits, of the most concise description (usually, the shortest program for a Universal Turing Machine or Bernoulli-Turing Machine[107]) of the physical entity or bit string for a given level of accuracy. This measure of complexity can easily be applied to bit strings by evaluating the question of whether the string can be *compressed* in such a way that the compressed bit string together with its required decompression procedure can be specified significantly more succinctly than the original string. That is, the complexity of the compression algorithm is irrelevant; we are concerned only with whether it is possible to compress the string in such a way that it can still be decompressed by a concisely specified algorithm.

Smith notes that in general the question of whether an arbitrary string is KCS random is undecidable. If we *do* know a way of compressing a string appropriately, we can say the string is *not* random, but if we don't know how to compress a string we can never say the string *is* random because there might still be some as yet undiscovered algorithm which could perform the feat. The observation is closely related to Gödel undecidability and to our previous discussion of computability.

We are interested in applying algorithmic complexity to binary codings of the behaviour of chaotic systems. Smith's example is an enormous simplification of the Lorenz model interpreted as a shift map over a bit string describing to which wing of the Lorenz attractor a given point will map at stroboscopically sampled times. (Note, incidentally, that this model ignores features of the real Lorenz system and is just the sort of dimension-reducing projection of complex dynamics onto a manifold which we criticised in the context of infinite intricacy.) That is, we start off the Lorenz model at a given point on the Lorenz attractor (for our purposes, we can read "on" as "arbitrarily near") and then measure it at periodic times to note the wing of the attractor to which the point has moved. This generates a binary expansion describing the gross behaviour of the trajectory on which the point lies, and we can apply the KCS measure of randomness to this binary expansion. Notice that the kind of

---

[107] A Bernoulli-Turing Machine is a Universal Turing Machine augmented with a random number generator.

"model of the model" which Smith has offered is deliberately designed to be a *very* simple model indeed. In the context where Smith uses this picture, the dynamics of the real Lorenz model become almost entirely irrelevant, since the binary specification of an initial condition in a single dimension simply describes with each succeeding digit which "wing" of a now irrelevant attractor a now irrelevant trajectory is visiting at each periodic sample. That is, the initial condition .1101110001 *just means* "wing 1, wing 1, wing 0, wing 1, wing 1, wing 1, wing 0, wing 0" and so on.

This is the point at which it becomes confusing to see what we are supposed to learn about the real Lorenz system from Smith's argument. He observes that most finite sequences count as random by the KCS definition. Indeed, it is interesting for us to note that in the infinite space of all possible *infinite* strings, not only is the measure of the set of nonrandom strings zero, but there is an *uncountably* infinite set of strings which are noncomputable and which thus cannot be generated by any algorithmic process, regardless of the length of the string we might use to specify the process if there were one. Thus even if every single string which could be recursively generated could also be generated by a finite program shorter than the string itself (an impossible proposition, of course), we could still say "most" strings were random (in the same sense in which there are infinitely more real numbers than integers). This is true even when we include the comparatively rare cryptoregular strings such as the binary expansion of π or of the square root of 2. But returning to the finite case, Smith notes that the proportion of sequences of a given length which cannot be specified by a program at least k bits shorter (i.e., the proportion of random sequences) must be greater than $1 - 2^{-k}$. (Of course this becomes more accurate the greater the string length we are considering.) Thus, the portion of sequences which can be compressed by more than 20 bits will be less than one in a million. He goes on to suggest, essentially, that since almost every binary sequence of any considerable length is KCS random, the behaviour of the Lorenz model understood as a shift map is almost always KCS random. In his own words,

> "an arbitrarily chosen 'seed' x_0 will, with a probability which can exponentially approach one, yield a sequence of n visits to the two sheets of the (simplified) Lorenz system which is as KCS random as almost every sequence of n coin tosses." (p. 67)

He is careful, incidentally, to point out that we are here concerned with a measure of the randomness of the *output* of the system, not of the actual *means* by which the output is produced. (Computer scientists would use the term "pseudorandom" to describe a string which appears random—and might even be KCS random—but which was created by purely deterministic means.) But of course, there is no real dynamical system at work here anyway; there is just straightforward string manipulation.

Moreover, recall that Smith is here discussing a *very* simple model of the Lorenz system. The model is so simple that the argument about coin tosses is almost trivial. Let's forget for a moment everything we know about the real Lorenz system. "Visits to the two sheets of the (simplified) Lorenz system" are nothing but the successive bits of arbitrarily chosen binary numbers, and we certainly don't need KCS or any other technical description of complexity to tell us the relationship between arbitrary bits and arbitrary coin tosses. What are we supposed to learn about the real Lorenz system from this argument to show that arbitrary 1s and 0s are just like arbitrary heads and tails?

At first we might think that we could learn very much from it if there were only a straightforward relationship between the real Lorenz system and the simple shift map—we might learn that in some sense the *output* of the Lorenz system is indistinguishable from random output. That is, maybe if we could find the right way to look at it, we could get behaviour from the real Lorenz system that was just like that of the simple shift map, and that behaviour would be indistinguishable from random. But of course, it's no secret that if we look at almost anything in just the right way, we'll get random-looking behaviour anyway. (See the next chapter for an example of finding "randomness" in a deterministic digital computer.) Moreover, to reiterate our warning when introducing this model of a model, Smith's modelling of the model is just the sort of endeavour which we criticised in the context of infinite intricacy, and it is just the sort of endeavour which we earlier saw was a source of trouble in understanding cognitive systems solely on the basis of ψ level dynamics.

This objection aside, it seems very problematic in any case to establish a relevant correlation between the two systems which allows us to keep Smith's desired conclusion on board. One problem is that if we strobe the real system at regular intervals, and if there exists any initial condition to give us a particular finite bit string, then there exists an

infinite class of different such initial conditions. This is a straightforward implication of the Shadowing Theorem: for any initial condition which does the trick, we are guaranteed a neighbourhood of other initial conditions which shadow the first closely enough also to do the trick. (We could prove a similar but weaker many to one correspondence by applying the pigeon hole principle: there is an infinite class of initial conditions but only a finite class of bit strings of a given length.) The problem is even worse, because the real system, unlike the model, needn't zip along from wing to wing at a constant speed. So for a given strobing rate, there could be a host of other trajectories which gave the same chosen string but which fitted in "extra" visits to other wings between sampling times. We could go on with other troubles, perhaps including stepping into a higher dimensional model of strobed output dynamics in the space of all possible outputs in order to show that in any given neighbourhood of the original system we can find some initial condition offering any desired strobed output, but the problems are already enough for the point we are making.

That point is simply that once we have established this many to one correlation by looking at the *real* Lorenz system, *there is no guarantee that strobed outputs will be uniformly distributed* in the space of all possible outputs. Once that guarantee is lost, there is no point to be made about similarity of output to random coin tosses. Smith's original observation about the shift map model was so correct as to be almost trivial, but without a significant amount of additional argument, which doesn't look too easy to provide, it just does not apply to the real model.

Unfortunately, Smith's final conclusions about the randomness of chaotic output appears to try to extrapolate the conclusion about the shift mp model to the kinds of models we use to try to explain the real world. He says,

> "The behaviour of a chaotic model is often equivalent to some variety of shift map defined over digit strings representing real numbers. In other words, the randomness we find over time in a trajectory's behaviour is the randomness that comes for free as we walk along the expansions of typical real numbers... But if that's so, when we turn to apply the model to the world, why shouldn't any discerned chaos just be an artefact of the modelling medium (simply chaos, as it were, dug out of the real number system used for building the model, rather than the world)?" (pp. 69-70)

If we take on board the first sentence, everything after is spot on. But the behaviour of typical chaotic models is *not* often equivalent to a shift map *until* we start strobing the system or otherwise reducing its dimensionality in some other way which obscures the *continuous nature of real systems.* (And again, it is no secret that if we look at most things in the right way, they can be pretty random-looking.) Real systems in Nature give us no indication that they interact only at stroboscopically sampled times or that they have dimensions which are irrelevant to a deterministic description of them. If we are interested in building a model of why the world looks the way it does when viewed through strobed glasses, that's one thing, but if we are philosophically interested in what actually "makes Nature go"—and this is the project in which I suggested in the previous chapter that we should be interested—that's another thing altogether. For that project, this line of thought is not relevant.

Having rid ourselves of any suspicion that this line is relevant to our own aims, let's explore some other questions about complexity as applied to the output of chaotic systems.

18.1.2 Chaotic Compression and KCS Variance

Smith notes (1993, p. 27) that it appears that every sequence of integers (such as 4, 2, 6, 13, 1, 99...) has a corresponding point or set of points on the Lorenz attractor which will take exactly the number of orbits around each wing of the attractor which is named in the sequence. So for our example, there is some initial condition which initiates a trajectory which travels around the left wing four times, the right one twice, the left one six times, the right thirteen, and so on. (Alternatively, there is a point which will generate any binary sequence.) Proving this is no small task, but it is an idea which I believe most mathematicians would find highly plausible. Indeed, it seems no more implausible than the idea that any region in phase space will eventually be visited by some point lying in any other region, which is just topological transitivity, one of the defining characteristics of chaotic systems. But if it is true, then it follows that chaotic systems such as the Lorenz model can be treated, in effect, as information compressors, exactly in those circumstances where the initial conditions generating a given sequence can be specified succinctly.

It may not be a computationally tractable or even a computable matter to determine one of the points in phase space which will lie on a trajectory following orbits describing an arbitrary sequence, and it certainly isn't the most efficient way of compressing a string for the general case, but this doesn't stop its being possible for there to be an initial condition at a computable point in phase space which will generate a description of any computable sequence whatsoever. Now, in some cases, specifying the initial point may take more bits than the desired string itself, and in some cases specifying the point together with instructions for generating the trajectory by applying the equations could also exceed the length of the string. But in other cases, there might be huge savings in length. For instance, every point on the Lorenz attractor specified by coordinates of a fixed length generates an infinite class of sequences of length from one to infinity; by applying other recursive functions of constant length to the output of the model in terms of numbers of orbits (perhaps even by feeding outputs back into the Lorenz system), we can generate an even broader class of similarly ordered such sequences.

Indeed, the idea of reducing some lengthy description to a more manageable size by coding the description in a chaotic system is the basis of the secretive commercial work of pioneering chaos researcher Michael Barnsley. By applying the so-called Collage Theorem (Barnsley 1988), it is possible to achieve extremely compact and progressively more accurate lossy representations of a pattern with a chaotic iterated function system. Similar methods may be involved in the image processing technique known as FITS (Functional Interpolating Transformation System), developed by Paris-based FITS Imaging for its software *Live Picture*. The software essentially translates a complex bit mapped image into a set of equations representing the same picture. The image can then be manipulated extremely rapidly because the huge amounts of data in the original image don't need to be altered, just the compact mathematical representation.

Given the very high compression ratios possible with this kind of method, we can easily see that *some* binary sequences produced by the Lorenz model will not be very KCS random at all if we can specify their initial conditions concisely enough. But there is no obvious relationship between the number of digits we might use to specify an initial condition and any qualitative "character" of the binary sequence the ensuing

trajectory will describe. Indeed, we shouldn't expect there to be because the behaviour of an individual trajectory depends on the *dynamics* of the system, *not* on the particular number system we use to specify that trajectory. Thus, the KCS definition may allow that two very similar (on some interpretation) binary strings produced by the Lorenz system could differ greatly in complexity according to whether the strings are compressible as initial conditions (specified with a limited number of bits) for the Lorenz system itself. In other words, for similar strings of a given length there may be nearly arbitrary variance of the KCS measure of complexity.

To use a drastic example, suppose it just happened to be possible to compress a bitmapped image of John Major by specifying an initial condition for the simplified Lorenz system which only requires five bits to express precisely. That is, if we gave the Lorenz system that initial condition, it could generate a pattern of bits which corresponded to those in our bitmapped image. But suppose a similar image of Douglas Hurd couldn't be generated by the Lorenz system unless we specified an initial condition with four thousand bits. (Here I don't mean we're using any greater precision: in this context *exactly* 2.11 is just as precise as *exactly* 2.247935769843459820498.) Ignoring for the moment the capabilities of other compression strategies—all of which, for all we know, might be substantially less efficient than the present method—we are left with the conclusion that John Major has a much more compressible (even soft and squishy!) image than Douglas Hurd. So Douglas Hurd is more complex, or algorithmically random, depending on the preferred interpretation. But this is a very counter-intuitive notion of complexity! Why should the output of the Lorenz model objectively *look* any different in terms of randomness or complexity when it started out at a point specified precisely with only five bits of information than when it started out at a point specified precisely with four thousand bits of information? (We might make similar arguments by appealing to other compression strategies, but it is enough to note that without reference to any compression standard at all except the system itself, we are left with a counterintuitive notion of randomness.)

With such observations in mind, we now move to an exploration of alternative measures of complexity which may help us to address these and other questions.

18.2 Alternative Measures

To start, the KCS definition of algorithmic complexity corresponds intuitively much more to randomness than complexity. Complexity generally evokes the notion of something that is highly ordered rather than something uniformly disordered. We won't debate the semantic point of whether we should consider uniformly disordered things complex or only grant that honour to things with hidden order; instead, I suggest simply that it is useful to consider this kind of difference between order and disorder and that such a consideration may shed some light on the question we have been considering about the relationship between the randomness of deterministic chaos and the randomness of noise. We might expect that a measure of complexity in the sense of the presence of order should often return "opposite" values compared to KCS complexity. We might hope that in adopting such a measure, we could also overcome some of the deficiencies of the KCS measure as discussed above.

18.2.1 Logical Depth

One such candidate measure is the logical depth of C.H. Bennett (1987, 1990) Logical depth is defined as the execution time of the shortest program for a universal computer (such as a Turing machine) which can generate a description of the object in question. More precisely, it is the harmonic mean of all such programs, since there may be a large class of programs of the same length which could generate any finite object.[108] The idea here is that logically deep objects (or binary sequences or whatever) should contain internal evidence of most plausibly having been the result of long computations or dynamical processes. (Note that wholly disordered strings can be generated quickly by long programs, whereas highly ordered strings might be generated more slowly by shorter programs.)

The most striking difference between logical depth and KCS as measures of complexity is that KCS returns a high complexity for both nondeterministically created disordered sequences and arbitrarily chosen sequences described by the Lorenz model, whereas Bennett's measure is

---

[108] This is analogous to the idea that the graphs of an infinite number of polynomials may pass through any finite set of points.

meant to give low complexity to very disordered and perhaps nondeterministically created strings and higher complexity to strings generated (or *generable*) by deterministic chaotic processes. Rather than simply the "opposite" value that we might have expected, logical depth hopes to discriminate between sequences essentially in terms of the complexity of how the sequences may be generated (i.e., execution time) rather than the "complexity" of the strings themselves.

An improvement on the logical depth measure, relatively minor but useful, has been offered by David Deutsch. (1985a) Deutsch's measure of quantum logical depth, or Q-logical depth, is keyed to the harmonic mean of the execution times of the shortest programs for his own Universal Quantum Computer. The key point of difference with ordinary logical depth stems from the suggestion that in Nature, random states are generated not by "long programs" but by short programs exploiting nondeterministic hardware. The quantum analogue of logical depth solves this minor problem by generating random sequences with very short programs. Yet, since we are primarily concerned with the execution times on particular machines, this change is unlikely to alter the measure of complexity by anything other than a uniform constant amount. Despite the theoretical elegance of Deutsch's quantum computer[109], it is difficult to see how the measure of complexity returned would differ markedly from that provided by logical depth measured with a Bernoulli-Turing Machine.[110]

However, a significant advantage of Q-logical depth is the ability to consider complexity across worlds in the Everett interpretation of quantum mechanics. The Everett interpretation is widely considered to be experimentally indistinguishable from other interpretations of quantum mechanics (Deutsch's own 1985b objection notwithstanding), and ordinarily I would not advocate reading too much into ways of thinking the interpretation seems to encourage. Yet in this case, I believe it is a useful tool for getting a handle on the kind of complexity we are discussing. In particular, we can interpret Q-logical depth as containing information about all universes (i.e., all states in the quantum linear superposition of the Universal Quantum Computer as it generates the

---

[109] Such computers are problematic. (Chapter 4 and Mulhauser 1995 in press)

[110] Also, it is very well to note that short programs for the Universal Quantum Computer can generate random strings rapidly, but if we are *given* a random string and asked to *supply* the Q-program to generate it, we are back where we were with Bennett's measure.

state in question) simultaneously (this means, incidentally, that the Q-logical depth is not an observable).[111] The Q-logical depth indicates high complexity only for objects which are present in all universes. As Deutsch puts it,

> "Observationally complex states that are different in different universes are not truly deep but just random. Since the Q-logical depth is a property of the quantum *state* (vector), a quantum subsystem need not necessarily have a well defined Q-logical depth (though often it will to a good degree of approximation). This is...to be expected since the knowledge in a system may reside entirely in its correlations with other systems." [emphasis original] (Deutsch 1985b, pp. 114-115)

For the moment, we are primarily concerned with the first sentence; I repeat the rest of the quotation for the complementarity with the earlier discussion about interactive decoherence and environmental correlations in complex quantum systems. While this is a rather speculative point, we might consider a sort of "staying power" of complexity and note a parallel between high Q-logical depth and some measure of structural stability which reaches down to the quantum level. That is, those objects with structural stability across universes—i.e., through possible linearly superposed states—are the most complex. This is another way of interpreting the difference between strings created by nondeterministic random noise and those deterministically created by chaotic dynamics: truly nondeterministic noise fluctuates without pattern across all universes, whereas chaotic patterns are more stable by virtue of their concealed order (i.e., their complexity which emerges even with Bennett's classical measure).

18.2.2 Logical Depth—Problems Solved?

We are now in a position to evaluate whether Q-logical depth overcomes the problems which led us to seek an alternative to KCS complexity. Recall that one concern was that KCS does not discriminate between strings created by entirely disordered nondeterministic processes (with tossing a fair coin as the paradigm example) and those created by

---

[111] We are of course *not* talking about the philosophers' possible worlds here, but rather about the different universes which can be interpreted as being home to the various possible states represented in the quantum linear superposition of the state vector.

ordered deterministic chaotic processes. Of course if it is Smith's aim to equate the randomness of chaos for practical purposes with the randomness of noise, then it is not surprising that he should have chosen this definition of complexity! Our other concern was the arbitrary variation in the KCS measure according to whether a given string could be compressed just by giving the initial condition and governing equations of a chaotic system which generated it. It seemed counterintuitive that two otherwise similar strings (such as bitmaps of the Prime Minister and the Foreign Secretary) should have different measures of complexity solely because one resulted from an initial condition specified in some number system to a greater number of places than the other.

As far as the first concern goes, the problem is meant to be solved by either logical depth or Q-logical depth. Both measures return a high complexity for strings which can be generated by succinctly specified ordered deterministic processes. Smith notes after his initial discussion of KCS complexity that deterministic chaos can *in principle* be distinguished from nondeterministic, disordered random noise by finding the right way to analyse it. Logical depth and Q-logical depth try to incorporate this feature into the measure itself and offer us a better way of distinguishing chaos from "mere noise". Having said that, however, there remains a problem in what we mean by "succinctly specified".

Suppose it takes c bits to specify the equations of the Lorenz system. Then *each* initial condition specifiable by x bits generates an infinite class of strings of length $y_k > x + c$ bits in length (one for each length), where $y_k = x + c + k$ and $k \in \{1, 2, 3...\}$, which are compressible by the present method by k bits. And of course, there is an infinite class of initial conditions specifiable by *some* x bit description. (In case this sounds odd, consider that it is no different than saying there is an infinite class of integers, all of which are of finite length.) This is all consistent with the fact that the probability of compressing an *arbitrary* string of length y (in the space of Lorenz outputs of length y) by this method (regardless of y) is less than 2^{1-c-k}. This is because there are 2^y strings of y length and less than or equal to $2^{x+1} - 2$ unique strings of y length generated by x or shorter length initial conditions. Thus the probability (tossing out that extra constant $- 2$ to make a cleaner answer without compromising the general point) is less than 2^{x+1-y}, or 2^{1-c} of finding an x or shorter length initial condition to generate a given y length string, or less than 2^{1-c-k} that we will be able to

compress the y length string by k bits or more. In other words, most strings of length y begin at initial conditions which require greater than x bits to specify and won't be compressible by this method. (And should Lorenz outputs be uniformly distributed in the space of all possible outputs, the bulk of them would also be KCS random, indicating that they could not be substantially compressed by any method.) Thus, these strings will have a low logical and Q-logical depth because the shortest program to generate them will just print them out. We can at least be happy, though, that for those strings which *are* compressible as initial conditions for a chaotic system, logical depth, unlike KCS, should be able to distinguish their origin in chaos rather than in nondeterministic randomness.

Returning to the first problem, our excursion into the complexity measure shopping market has been even less fruitful. Because it is still the case that a sequence requiring an extremely precise description of the initial conditions of the dynamical system which created it could turn out as random as coin tosses on the logical depth measures, there could still be arbitrary differences in the complexity of similar strings, where one could be generated from a very succinct initial condition and the other only from an initial condition specified with very many bits. (Of course, this argument does not necessarily generalise over all possible compression algorithms; we are concerned for the moment just with the complexity of the string with reference to its own system's standard of compression in terms of specifying an initial condition.)

This second concern especially, then, appears to be a shortcoming of all the complexity measures we've tried so far. In an earlier version of this material, I suggested that it would be helpful to have a measure which didn't care about particulars of the starting conditions for a system, which commented only on the overall dynamics of the system itself (as opposed to individual strings produced by it). Although I was unable to suggest an alternative at the time, I would like now to offer one.

18.2.3 Functional Logical Depth—Problems Solved

I suggest a new measure which, by concentrating on the complexity of the *relationship* between inputs and outputs rather than on outputs or even processes themselves, hopefully allows us to avoid some of the pitfalls of the other measures while allowing us to describe new observations which may have eluded us before. By quantifying the

complexity of the relationship between a set of inputs to some "black box" system and the corresponding outputs, the new measure allows us to compare different systems in a way largely independent of our knowledge of the processes at work in the "black box". The new measure we will call *functional logical depth*, or F-logical depth, and it is defined as the mean execution time of the shortest Universal Bernoulli-Turing machine description of the input/output relationship of the system in question. That is, it is the average over all possible inputs of the length of time taken by the shortest program to produce an output identical to that which the system in question would produce in response to the same inputs. I hope mathematicians everywhere will excuse the silliness of trying to express this in the following way, but in the simplest two dimensional case, the functional logical depth F of a relationship described by Bernoulli-Turing program S for a given level of precision P (i.e., how many bits we're looking for in the outputs), where "pseudo-function" E returns the execution time to produce an output of precision P for a single initial condition specified in x and y, might look something like:[112]

$$F(S, P) = \frac{\int_{y_0}^{y_b} \int_{x_0}^{x_a} E(S(x, y), P)\,dx\,dy}{(x_a - x_0)(y_b - y_0)}$$

Here integration is over all the relevant x and y in the domain of the system, and the result of the integration is divided by the area of the domain space in order to give an average execution time. We can think of this graphically as a process of scanning across a plane of possible (x, y) and plotting a height along an E axis corresponding to the execution time to produce the output for each (x, y). The two integrals give the volume under this E manifold, which we then divide by the area of the (x, y) domain to derive the average height of the manifold above the (x, y) domain. Notice that F-logical depth is closely related to but not identical with the computational complexity of algorithm analysis, in which

112 This is really simplified beyond plausibility, with, for instance, an easy (x, y) domain over which we can integrate in this simple manner. Indeed, for some systems, both the (x, y) domain and the manifold in E might be fractal. I've deliberately written it out in this entirely inaccurate way purely for illustrative purposes; shortly in the text we observe that F-logical depth is not a computable value anyway.

algorithms are ordered according to the number of steps which must be performed in their completion.

A few minor points deserve attention before we move on to more significant observations about the measure. First, for a nondeterministic system, we aren't concerned that S produce outputs identical with that of the real black box system—how could it?—but only that its probability density function matches to within P.[113] Second, in comparing outputs, we won't be concerned with the amount of time it might take otherwise output-identical systems to give their outputs, since for the purposes of F-logical depth, we want to measure the complexity relationship between inputs and outputs, not whether this system or that is quicker or slower or even incorporates some arbitrary time delay. Also, F-logical depth does not necessarily measure the complexity of a particular process, but only the complexity of the input/output relationship of that process. Thus, any two processes which instantiate the same input/output relationship are equivalent in terms of F-logical depth, regardless of their individual computational complexities. Finally, perhaps the clumsiness of the pseudo calculus will be excused in light of the fact that F-logical depth is no more a computable measure than KCS, logical depth, or Q-logical depth.[114]

There are several things more worthwhile noticing about F-logical depth as a measure of complexity. First, the value is at a minimum for nondeterministically random or trivial outputs. If there is no probability relation whatsoever between inputs and outputs, then for each input S can simply offer a scaled output from its random number generator and be done with it. Likewise, the F-logical depth will be minimal if the relationship is trivial, since S might either just look up the appropriate output value in a table or output a constant value or even output the input string—as in the case of Smith's shift map rendition of the Lorenz system—according to the particulars of the trivial system in question. (This observation also reveals that F-logical depth is more meaningful for systems with continuous or with discrete but very large input spaces. For

---

113 The usefulness of F-logical depth distinctions depends on the value of P. At low precision, all processes look F-logically shallow.

114 We might measure a quantity similar to F-logical depth by applying the algorithmic complexity measure to a bit description of the relationship between a system's inputs and outputs. This approach does, however, entail certain difficulties which I believe make F-logical depth preferable.

systems with discrete and small input spaces, it may always be quicker and shorter to exploit a simple lookup table than to undertake calculations more closely related to the behaviour of the real input-output system under consideration.)

At the other end of the complexity spectrum are the kinds of deep processes the products of which logical depth was originally intended to pick out. By defining the F-logical depth measure over all possible inputs, we have avoided one difficulty which arose for ordinary logical depth, namely, the problem of giving a low complexity to those outputs which truly were created by long and complex processes but whose initial conditions in terms of those processes required very many bits to specify. Again by focussing on an overall input/output relationship, this measure avoids the kind of variation in complexity which occurs when, for instance, offering a million generations of a million chimpanzees typewriters really does result in the creation of a highly ordered encyclopaedia. F-logical depth just indicates the process itself is of low complexity and makes no claims about individual outputs. Likewise, F-logical depth cannot comment on the relative complexity of the images of the British Prime Minister or the Foreign Secretary, but it can comment on the kinds of processes which lead to them and what relationship these processes may have to those of a million generations of a million chimpanzees.

To be sure, it is useful to be able to comment on individual outputs, but this shouldn't be seen as a criticism of F-logical depth: it isn't intended as a replacement for the other measures, but simply as a complement, offering something the others do not. With the new measure in hand, we may now examine the last of the conclusions Smith would like to make which may be relevant to our own project.

18.3 Chaotic Randomness and Random Randomness

The first and simplest conclusion to note is that in terms of the new measure, we can see clearly that the randomness of deterministic chaos does not equal the randomness of nondeterministic noise. Let's be very clear on this point: the three properties to which we have appealed to characterise chaos are input/output relationships. Outputs are sensitively dependent on inputs (SIC). The set of possible outputs which repeat is

infinite, and in the input space, the set of inputs which give these outputs is dense (dense covering of periodic points). Finally, some possible input arbitrarily similar to a given one will eventually give an output which is arbitrarily similar to some other given one (topological transitivity). Thus, we are perfectly justified in applying *this* measure as the standard by which to make comparisons of complexity between *systems*. And once more, the conclusion that chaotic systems can, if we look at them in the right way, give *outputs* which look random on the other measures is singularly unremarkable. But what about the relationship between F-complex chaos and the noise which Smith emphasises is part of the boundary conditions of every system?

18.4 Noisy Chaos

Let's turn to the final conclusions Smith wants to draw from his application of the KCS measure of complexity to the output of chaotic systems. Smith notes as we have said before that it is possible in principle to distinguish the "randomness" of deterministic chaos from that of nondeterministic noise—and the F-logical depth measure gives us the conceptual framework to do just that.[115] But, he says, the randomness (by the KCS definition) we observe in chaotic systems is the kind that "comes for free" as a result of the system's nonlinearity extracting detail about the real number specifications of the system's initial conditions. But this kind of detail, he argued previously, is incompatible with the kinds of abstractions to which we typically apply chaotic models. Thus perhaps physical phenomena shouldn't really be modelled with the infinitely detailed real number system. Instead, perhaps we need noisy models of limited precision which reflect Smith's belief that all physical systems are subject to low level random noise (strictly nondeterministic noise, we might wonder?). Thus, he cheerfully finishes, what really matters is the qualitative behaviour of the limited precision models we run on digital

---

115 Of course, we can also draw the distinction if we have empirically discovered some of the underlying order of deterministic chaos, but F-logical depth offers a framework when this empirical data is lacking. It is important in this context not to slip into the non-technical definition of chaos—meaning utter disorder or whatever—and then draw some conclusions with respect to its relationship to noise, eventually foisting that conclusion back onto the shoulders of chaos in the technical sense.

computers, subject as they are to continual noise in the form of roundoff errors, or limited precision computations.

Now, we have made the point in several places before that odd things may happen when we artificially reduce the dimensionality of a system of which we are wanting an explanation, and this observation comes into play here again. We won't reiterate the response to Smith's points on abstractions and what level of precision is appropriate for them, as this is by now well trodden territory. But now we can make similar points about noise.

18.4.1 Who is Making Noise?

The first observation we may make is that there is something curious about the assertion that all physical systems are subject to continual low level random noise. If Smith means noise that is random on the KCS definition, then this could be either deterministically produced chaotic "noise", or it could be truly nondeterministic. If it is the former, then Smith cannot appeal to this point as an argument against modelling in the real number system, nor can he appeal to it as an argument for the kind of "noise" created by roundoff errors unless he is prepared to offer an argument that roundoff errors are chaotic. In fairness, he does not make any direct appeal to this first argument against real numbers, although it would be easy to interpret Smith's final comments as mutually supporting assertions painting a particular picture of reality, rather than as a well structured linear argument. If, on the other hand, Smith means noise that is random and nondeterministic, then he is sidestepping a subtle question and leaving us with the impression of a powerful ontological position for which he offers no argumentative support, namely that there exists nondeterministic noise in the world which does actively influence every possible real physical system.

To take this impression first, as physicists or engineers, we of course find it useful to model systems with environmental noise rather than trying to track the evolution of every single particle which could have an influence on the system being modelled. But as philosophers, we remain interested in how a system "really" works—what makes a system go—in the spirit of our earlier comments on realism. It is fine to assert that what really matters for grasping the qualitative behaviour of physical systems is the kind of simulation we can run on a digital computer. But it is

philosophically unsatisfactory just to *assume* that truly nondeterministic noise exists everywhere and influences every conceivable physical system (and that this noise is qualitatively similar to computer roundoff error). This is particularly true considering the availability of coherent entirely deterministic interpretations of quantum mechanics. (Bohm 1952) The perturbations of quantum vacuum theory (see Podolny 1986 for a charming introduction and a romantic history of science in the former Soviet Union; also Puthoff 1989, 1990) may be the best candidate for a real nondeterministic noisy background of the sort Smith may be proposing, but as far as I know there is no *a priori* reason why it, too, cannot be subsumed under a comprehensive deterministic but nonlocal interpretation of quantum physics. Indeed, all-pervading nondeterministic background noise is arguably incompatible with the very project of quantum cosmology. If truly nondeterministic noise does not necessarily come from quantum mechanics, then where does it come from?

18.4.2 Levels of Description Ad Nauseum

The answer, I believe, lies in the question which Smith's approach has sidestepped. That question concerns whether noise is random relative to the system being modelled or random relative to "everything", as it were. In other words, is noise, instead of being some all-pervading background hum, actually an entirely deterministic and perhaps wholly ordered (and even F-complex) effect of some other system, which gets "sampled" by our model just as Smith's model of the model "samples" the Lorenz system and winds up with mostly KCS random output?

There is an interesting parallel here with the discussion of interactive decoherence in complex quantum systems. As we saw earlier, until they decided to analyse quantum systems at a more detailed level of complexity, physicists were saddled with the idea that consciousness was essential to quantum measurement. It was only when more and more of the dimensions actually relevant to a quantum system were considered that this problem disappeared. Likewise, it may be that when quasi-classical systems are modelled in greater detail, "random noise" reveals itself as an entirely deterministic but perhaps chaotic lower level influence. Although this point is clear within the F-logical depth framework, it makes little sense if we limit ourselves to KCS complexity.

It is possible to have an entirely deterministic model of a physical system, or it is possible to posit the existence of an all-pervading nondeterministic noise. KCS complexity makes it impossible to distinguish. Once again, if our project is to get hold of an ontological understanding of what really "makes Nature go" instead of only achieving the (perhaps very difficult) task of creating models that make good engineering sense, then we cannot simply sit back on our low-dimensional, limited precision laurels and say noise is everywhere.

18.4.3 Out of Our Depth?

One final observation about Smith's closing comments is that they preclude the kind of theoretical distinction offered by the eminent mathematician Steven Wolfram (1985) between *homoplectic* and *autoplectic* processes. Homoplectic processes, Wolfram suggests, result from those dynamical systems which generate macroscopic KCS random behaviour by magnifying the significance of environmental noise (which, on our interpretation, might really be the influence of chaotic systems otherwise external to the system being modelled). The admittedly speculative brand of autoplectic processes, on the other hand, would generate the same macroscopically pseudorandom behaviour (i.e. behaviour which may be KCS random but is complex on the F-logical depth measure) independently of the presence of noise. Such a robust autoplectic system could generate logical depth and maintain internal evidence of a long history, whereas homoplectic processes would remain comparatively shallow because of the randomising influence of noise. *If* there is nondeterministic Smith style noise, this kind of robustness is just the kind we expect to single out when we consider the multiple worlds view of Q-logical depth or a quantum analogue of F-logical depth. (Yet Smith's picture precludes it!) An autoplectic system would remain stable over possible universes, whereas the complexity of a homoplectic system would be wiped out by the "random" variance across universes (here it is useful to appeal to the vacuum fluctuations mentioned above).

Taking Wolfram's distinction as a point of departure, Bennett wonders interestingly if dissipative processes such as turbulence, which are not explicitly computational, could still generate logical depth. He wonders if something like a waterfall could be an autoplectic process which contains objective evidence of a long dynamical history: is there

any objective difference between a day old waterfall and a year old one? He does not answer the question, although he does cite Ahlers and Walden (1980) for evidence of "fairly long-term pseudorandom behaviour near the onset of convective turbulence". (Bennett 1990, p. 147) The question, however, is one which we must leave for another time.

18.5 From Complex to Simple

In any case, we have seen that KCS is not the only description of complexity on offer, and we have seen that Smith's argument for counting the outputs of chaotic systems as random is an unremarkable comment on the uniform distribution of real numbers. It is applicable to the real world only insofar as we can establish a useful relationship between continuous (differential) models and strobed or otherwise discretized (difference) models of lower dimension (politically correct term: 'dimensionally challenged'?) We have seen some of the advantages of using logical or quantum logical depth as our measure of complexity, although we have noted that they don't succeed in picking out all the deterministically created but KCS random strings. They also share a shortcoming of the KCS description in that arbitrary, otherwise similar, strings may receive different measures of complexity for no other reason than that they can be described succinctly as initial conditions of chaotic dynamical systems. These problems are overcome by functional logical depth, an alternative measure of the complexity of the *relationship* between inputs and outputs of a system, whatever the internal details of the system might be.

We have made a fairly critical exploration of Smith's closing conclusions about chaos, noise, and limited precision simulations and seen that there is much more to be said about the problem than is at first apparent using only the KCS description. It is time now to turn to a general class of problems which arises for our kinds of purposes when we try to use any of the standard measures of complexity to describe dynamical systems. Our examination of this problem returns us to the central theme of representation which was so crucial in our early understanding of self models, and offers us a glimpse into what, if anything, we should really be concluding about the possible characteristic of material substrates for self models which we have for so long been discussing, chaotic dynamics.

19

Complexity and Representation

19.1 Why Representation?

Consider some of the following representations of what is arguably the same string:

Representation 1:

0100001001101100011001010111001101110011011001010110010000100000011000
0010111001001100101001000000111010001101000011001010010000001110000011
1001010110000101100011011001010110110101100001011010110110010101110011
0011100110011101000100000001000000110011001101111011100100010000001111
0100011010000110010101111001001000000111001101101000011000010110110000
1101100001000000110001001100101001000000110001101100001011011000110110
0001100101011001000010000001110100011010000110010100100000011000110110
0100001101100011001000111001001100101011011100010000001101111011001100
0010000001000111011011110110010000101110

Representation 2:

426C6573736564206172652074686520706563656D616B6572733A2020666F722
0746865792073686616C6C2062652063616C6C65642074686520368696C6472656E
206F6620476F642E

These first two strings make little sense on first glance; the third representation is for most people most recognisable. The first is the same text written in binary (base 2) ASCII equivalent, where each 8 bit byte corresponds to a particular character, and the second is the same in hexadecimal (base 16) form, where each pair of characters indicates a byte. The fourth representation is a simple kind of code where entire words are indicated by single digits, and the final very economical representation is the book and chapter and verse from the King James Bible. A brief

consideration of these five ways of saying the same thing reveals that complexity is not a straightforward measurement and that we must understand the relationship between complexity and representation in order to make any sense of complexity as applied to philosophical questions about minds.

Representation 3:

Blessed are the peacemakers: for they shall be called the children of God.

Representation 4:

123456789ABCDEF where

1 = Blessed; 2 = are; 3 = the; 4 = peacemakers; 5 = :; 6 = for; 7 = they; 8 = shall; 9 = be; A = called; B = the; C = children; D = of; E = God; F = .

Representation 5:

Matt 5:9

19.2 Complexity in Representations

If we type the characters of representation 3 into a modern personal computer, they will be stored in the machine as simple binary switches from which we could extract exactly the bits in representation 1. When the computer actually fetches these bits from memory, it always does so in groups of multiples of 8 bits, often 32 bit chunks sometimes called words or double words. These 32 bit chunks correspond to sets of 8 characters in representation 2. Thus the first 32 bits of representation 1, 01000010011011000110010101110011, correspond to the first 8 hexadecimal characters 426C6573. Although everything in a digital computer is stored in binary form, computer scientists often use the simpler hexadecimal representation when analysing a machine or when writing low level code because it is shorter and faster to understand. Most people find it more natural to think in terms of hexadecimal representations than in binary expansions of the same thing.

But most people find it even easier to understand the text of the third representation, and in our everyday human to human communications this is the form we use. For our purposes, we might very loosely call this the *native* representation. In order to understand what is being said by either of the first two representations, we must understand how the bit patterns or hexadecimal patterns *correspond* to the characters of our ordinary text native representation. This correspondence is given by the ASCII table, the American Standard Code for Information Interchange. Without the standardised code, we would have no way of understanding the first two representations.

We can say something similar about the last two representations. Here, in order to understand the string 123456789ABCDEF, we must know the special table given above. Rather than representing individual characters with a special code, as in ASCII, here we are representing entire words. Likewise in the last representation, we are denoting an entire sentence by a simple verse reference. If we know what the King James Bible is, then we can understand Matt 5:9 just as well as we can understand 123456789ABCDEF if we know the special table.

The point of all this is that whenever we analyse the complexity of a string, we will *almost always* involve some representational scheme. Even when we considered the behaviour of the simplified Lorenz model and the bit strings it generated, we began with the representational scheme which said "when the trajectory is on this wing of the attractor, we'll call it a 1, and when it is on this other wing, we'll call it a 0". (The only case in which a representational scheme *won't* be involved is if we are analysing output in a system's native representation; more on this presently.)

But each representational scheme includes its *own* degree of order or complexity. For example, we could generate the string in representation 4 very simply by just writing down the highly ordered list of the first 15 positive hexadecimal numbers. But if the table defining the representational scheme had been different, such that the string were maybe C326B5F847D1E9A, generating the string could be much more difficult. By a simple change of the representational scheme, the string has gone from highly ordered to (apparently) not very ordered at all. But yet we are considering, in a way, the same basic thing: the string from representation 3. Likewise, ASCII representations of plain text like the first two above can be easily compressed by standard methods by about

50%. Thus, they have some pattern and complexity. Yet if we run the same strings through a few rounds of a randomising algorithm such as DES, the U.S. government's Data Encryption Standard, typical strings may be nearly impossible to compress.[116] Where does the complexity go? In a clear sense, it is present in the data encryption algorithm, which has essentially replaced the standard ASCII representation system with a more complicated one. Indeed, a change of the representational scheme can transform a description of almost anything into a description of almost anything else—a sort of information theoretic alchemy. We could devise some decoding scheme, for instance, where we could extract the first few bars of Beethoven's Fifth Symphony or the first few bits of a digitised photo of John Major from any of these five example strings. In general, there is some decoding scheme to find anything in anything.

Thus different ways of representing particular things such as English sentences or descriptions of Beethoven's music carry with them different consequences for the complexity measures. Insofar as discussing the complexity of a system almost always involves the use of *some* representational system, attributing complexity almost always involves a prejudgement inherent in the representational system we choose. Some representational schemes produce very dense and uncompressible strings, where superfluous information is stripped out by the complexity of the scheme, whereas some produce strings with plenty of inessential data which can be removed by compression. Indeed, particular representational schemes will give preference to particular kinds of order within that which is being represented. For instance, the kind of representation used in the last example above plainly gives preference to sentences which happen to be verses of the Bible. It is still an ordinary text representation, but it relies for its brevity upon the vast information stored in the correlations between short book names and verse indices to give a very dense representation.[117]

---

[116] Someone might object that DES has actually increased the logical depth of the string which was randomised, since it is now a more computationally intensive task to extract the original string and compress *it*, rather than just trying to compress the DESed string. But of course doing this requires a specification of the *entire* DESed string as well as of the required decryption algorithm; thus, the shortest program which can generate the DESed string likely will be one which simply lists its elements, and the string will be both KCS random and logically shallow.

[117] It's interesting to notice that when we compress a given string we are offering a different representation of the same data and including within the product string a

We've established that the same thing represented two different ways may appear more or less complex according to the characteristics of the representational system. One example of this is the Bible passage represented as binary ASCII as opposed to binary ASCII which has been randomised by DES. In a way, both patterns represent the same thing under different interpretations, yet the first appears logically deep because it is compressible while the other looks both KCS random and logically shallow. Moving from the representational scheme of ordinary ASCII to ASCII with a few rounds of DES thoroughly alters the complexity of the "same" string under both KCS and the logical depth measures. (Notice the high complexity of DES itself under the F-logical depth measure, by the way.) With the observation to hand that the choice of representational scheme largely determines the outcome of a complexity measure applied to a *system*, let's consider what it would mean to measure complexity in a system's "native representation".

19.3 Complexity Without Representation

The easy example here is analysing the internal functions of a digital computer, since the complexity measures we've considered are all meant for application to bit strings, and that's exactly how data are represented internally in a digital computer. Since monitoring bit streams at some points would yield highly ordered patterns and at others highly disordered patterns, we'll consider two examples. First, suppose we watched the signals across control lines from a floppy disk controller to a disk drive. The fact that there is a limited number of possible meaningful signals which the controller can send to the disk drive (turn on the main motor, strobe a stepper motor switch or a data latch, reset the read/write head, etc.) means that only particular patterns will ever occur. There will also be a higher level of order in that particular commands frequently follow each other (such as turn on the main motor and strobe a stepper motor) while others never do (such as strobe a stepper motor and then turn on the main motor). Note that these individual patterns *themselves* might be uncompressible, but the order present in the way in which they

---

definition of the new representational scheme in terms of the original one: a key for one-way translating, or decompressing.

are used will give us an easy way of compressing the overall pattern.[118] Thus *on an appropriate scale* the stream of bits we might read across these lines will look highly ordered and easily compressible on any of the three measures of complexity we've considered. The immediate lesson of this example is that patterns which look from an external point of view to be highly disordered or random (but recurring) might be put to very ordered functionally relevant use in a digital computer. While the individual patterns might be highly disordered, *for the computer* they represent instructions in a well ordered process.

Now, in the same digital computer system, we could also measure the signals along functionally relevant lines where the patterns would not be nearly so regular. For instance, in a computer with 9-bit parity checked RAM, we might decide to sample the line carrying the parity bit. Computers with parity checked RAM store an extra bit along with every 8 bits of data to indicate whether there is an odd or an even number of high bits (i.e., 1s) in that memory location (or, equivalently, to make the total number of 1s either even or odd). While there will be localised exceptions for programs which frequently access highly redundant sequences of bytes in memory, in general if we watch a computer for long enough while it executes some code the value on the parity line of the computer's data bus will be patternless, indicating a 1 about as frequently as a 0.[119] Under the measures of complexity we've considered, the strings we sampled along this line would be considerably more random and uncompressible than the strings we measured along the disk control line. The uninitiated philosopher of digital computers might even conclude that the value on this parity line constituted a computer equivalent of nondeterministic noise. The immediate lesson from this example is that even in a highly ordered system such as memory accessing in a digital computer, appropriate measurements may yield strings with little or no apparent pattern.

Now we have two examples of sampled signals from a paradigmatically well ordered digital computer, and one is highly compressible while the other appears fairly random. Yet they are both

---

118 The particulars of individual signal patterns depend on the instruction set of the relevant chips, of course, but there is nothing in digital computer design which prevents us in principle from using the most "random" strings possible to represent whatever we want.

119 This is part of the theory of parity checking: the reliability of the scheme depends on an unbiased, fairly uniform distribution of evens and odds.

from functionally relevant areas of a system which at the level of description we are considering is entirely deterministic and completely free from noise. And since it is not necessary to translate the quantities we're measuring in any way, we needn't worry that our choice of representational scheme has influenced the complexity result in any way. We have noticed that highly disordered patterns might be put to highly ordered uses and that likewise highly ordered processes might generate highly disordered patterns.

Our entire discussion has been motivated by questions about the complexity of chaotic signals and their relationship to nondeterministic noise, specifically as these questions bear on the presence of chaotic processes in the brain. After exploring the relationship between representation and complexity as well as asking particular questions about complexity in native representation in the digital computer example, we are now in a position to say something about what complexity measures might say about chaos in the brain.

19.4 Complexity in Natural Chaotic Systems

First lets consider the easier question: if we sample signals in the brain, convert them into binary patterns, and analyse their complexity, what conclusions might we draw? Well, if we happen to sample in an area where signals look chaotic (as Freeman, for instance, has done), we might find a KCS random pattern which could also be a logically deep one. Regardless of whether the pattern actually was logically deep, it might be akin to the individual command patterns which the disk controller sends to the disk drive. That is, the chaotic pattern could itself play a functionally relevant rôle in a process ordered at another level. To use an example inspired by Freeman, the chaotic pattern might simply be the firing pattern of neurons in an area of the olfactory cortex which is presently sensing the odour of bunny food. Just as the individual disk drive controller signals may be patternless yet represent *for the computer* functionally relevant commands in a well ordered process, the chaotic firing patterns in the rabbit's cortex may represent *for the rabbit* a piece of data functionally relevant to the process of getting a bite to eat.

Moreover, despite the fact that detectable order might not emerge until signals are analysed at a higher functional level, we cannot even

conclude that the special properties of chaos such as sensitive dependence on initial conditions are irrelevant at the lower level. (Something like this we *could* do in the disk controller example, where relationships between bits in a particular disk drive command have no functional rôle at *that* level, although they do obviously have a crucial rôle at the level of the logic gates of the chips activated by them.) It might be the case, for instance, that something like SIC is a property required by neural systems capable of making the kinds of fine discriminations displayed by the olfactory cortex. While the higher level order may be non-chaotic, the lower level order may be *essentially* chaotic, ordered on the functional logical depth measure but perhaps not on KCS.

Alternatively, finding an apparently chaotic signal, or even a signal which was random on *both* the KCS and logical depth measures, could be akin to finding apparently patternless signals on a parity line. We might be measuring an apparently random (epiphenomenal) byproduct of a functionally relevant, logically deep process. Likewise, whatever we did find might be merely the result of our choice of representational scheme. These possibilities for interpreting chaotic or even random (on the logical depth measures, as opposed to just KCS) signals in the brain suggest that without much more information there is very little we can conclude about the rôle the signal plays in the brain's functioning.

This brings us to the more difficult question about complexity measures applied to brain functioning: how can we move from an analysis of complexity of a representation of brain functioning (such as binary representations of spiking frequencies) to an analysis of complexity in the brain's "native representation"? (In particular, how can we begin to draw conclusions about the presence of deterministic chaos in the brain and how this could be related to nondeterministic noise?) This is akin to moving from an analysis of disordered individual disk controller commands to the functionally relevant and ordered use to which these commands are put. It is akin to moving from the level of describing the less ordered precise positions in 3-space of particles of toner on a white sheet of paper to the more ordered level of the actual text being represented by those particles of toner.

19.5 Chaos and Complexity in Neural Systems

Answering this question is by no means easy, and it is one on which we shall here make only rudimentary headway. Some people might like to say that there can't really be functionally relevant chaos in the brain because otherwise people's behaviour would be chaotic. Others, like Smith, appeal to the KCS definition of complexity to put chaos on the same footing as nondeterministic noise; on this view, it is hard to see how chaos in the brain could play any important rôle at all. But both these naïve approaches are unable to accommodate the kinds of observations we have made. They do not begin to answer the question of how to analyse chaos in the brain in it's native representation. So let's take note of some very general points that can be made.

First, it is possible that all chaos in the brain, if analysed in terms of the uses to which it is put by the brain itself, would be something like the patternless codes which might be sent to a disk drive. In other words, we could theoretically change the chaotic pattern to any other pattern, and as long as we preserved the functional rôle of the new pattern with respect to the old, nothing would change. Even if it weren't physically possible to achieve such a change with a real biological neural system, perhaps because neurons are structurally disposed to produce chaotic spiking behaviour, it could still be true that any artificial rendition of an intelligent system could get by with a nonchaotic pattern in place of the chaotic behaviour of the real brain. Moreover, in this case, a chaotic pattern could just as easily be wholly nondeterministic noise (as long as bits of such noise were reproducible).

Alternatively, all chaos in the brain might turn out to be nothing but a byproduct of interactions between separate nonchaotic subsystems. In this case, it might be impossible to achieve the same kind of information processing with neurons without creating the same chaotic byproduct. Along similar lines, it might be impossible to build a wood fire for cooking food without producing smoke, but the special properties of the smoke might be entirely irrelevant to the cooking process. If the chaos were a result of peculiarities of biological neurons but not of simulated intelligences (in the sense that smoke is a property of burning wood but not of an electric hob), it might be possible to implement the relevant

information processing artificially without creating the byproduct chaos (just as it is possible to cook food with a smokeless electric hob).

Finally, it could be that the special properties of chaotic systems are functionally essential at the level where they are present. We discussed in Chapter 13 some of the possible applications of chaos in cognitive systems and some of the capabilities which its presence might help explain. We suggested, for instance, that SIC might be important for any neural network capable of recognising fine differences between odours. Such sensitivity could also be important for sensing changes in the position of a motor control system, and topological transitivity might be important for rapidly "searching" wide neighbourhoods of an area of phase space. SIC and dense coverings of phase space with periodic points could be important for the creativity of a system able to make nonlogical "leaps". Yet all these capabilities might at the same time be parts of a process which was essentially nonchaotic at a higher level. None of these capabilities—recognising fine differences in odour, sensing positions for motor control, searching large spaces, or being creative—imply in any way that higher level processes which appeal to them will themselves be chaotic. This takes care of the concern stated above that real biological systems can't be relevantly chaotic because then their higher level behaviour would be chaotic: a chaotic process may simply provide input to a higher level process which was not noticeably chaotic. (But recall the literature we discussed in Chapter 14 which indicated ways in which chaos may persist into high level brain dynamics.)

19.6 Complexity's Last Representation

In the end, of course, conclusions about the rôle of chaos in biological intelligent systems and its relationship to nondeterministic noise will have to wait for more empirical data. It is primarily a job not for philosophers but for experimentalists. Philosophers have an important part to play in analysing the possibilities and developing testable theories which can guide experimental explorations, but they have no place in prejudging the whole question by appealing to some abstract measure of complexity. The importance of chaos theory cannot be ushered away by the generally appropriate but overzealous naysaying of Smith, but likewise it is not the magical solution to all the problems of

consciousness that it sometimes seems trumped up to be in the popular press (and even by some otherwise sober-minded philosophers!).

We have come a long way from the original view that chaos was essentially little different from noise and that all that really mattered in terms of analysing real physical systems were the kinds of limited precision simulations we can run on digital computers. The point of the previous few chapters has been to explore some properties of chaos and their possible relevance to intelligent systems and then to defend our points against the temptation to devalue chaos to irrelevance. Chaos as it relates to questions of mind is an area ripe for more experimental exploration and theoretical guidance, and it cannot be dismissed so easily.

20

Tell Them What You've Told Them

We noted in the Introduction that our strategy in this dissertation is akin to that of the physicist who is trying better to understand the weak nuclear force. Our physicist takes an educated guess that perhaps things called *w*-particles exist, and she goes on to see what sorts of things she could explain with *w*-particles. She investigates some of the consequences of positing *w*-particles, she asks whether *w*-particles are plausible in the context of her existing hypotheses, and she explores some of the empirical and theoretical data with which it is most important to see if *w*-particles are consistent. As her research programme proceeds, she comes to a better understanding of the use of the *w*-particles hypothesis, and hopefully she comes to a better understanding of the weak force.

Likewise, the bulk of this dissertation has been dedicated to sketching the general plausibility of a number of different positions on fundamental issues in philosophy and investigating what problems those positions might help solve, what other positions they are compatible with, and what empirical and theoretical data they need to be checked against. It is now time to look back over the progress we have made and take stock of the usefulness of our original and most basic positions.

20.1 Under the Influence

Let's start with the two underlying themes for which we never explicitly argued but which have influenced much of our analysis of more major positions: the concern for different levels of description and points of view. With respect to the first, we've seen repeatedly the importance of examining systems at more than one level; often substantive conclusions have emerged when we've considered data drawn from different levels of description of the same system. There's more than one way to describe a

cat (or a self model)! In our discussions in the second half of the dissertation, we've seen how dynamics at one level may profoundly influence the possible behaviour of a system at a higher level. Likewise, we saw in the first half some ways in which an explanation proposed at a given level to explain a particular phenomenon—such as our account of the emergence of conscious sensation from a material substrate—may demand certain characteristics of dynamics at a lower level.

Our awareness of different points of view has been similarly helpful. In particular, the *complementarity* of the first person and third person points of view underlies our cybernetic realism position. In the case of levels of description, we began with a whole unitary system and then deliberately introduced different aspects of it by looking to characteristics which appear at particular scales. In the case of points of view, we began with two distinct aspects and attempted to bring them together into a whole. Especially with our discussions of what we've dubbed the first person problem and the third person problem, we've taken some steps toward achieving this goal of unification. While the two points of view are important in their own right, just as descriptions at different levels are important in their own right, we've seen they are also important as different aspects of the same thing. Now let's examine whether our more explicit positions have served us as well as the underlying influences.

20.2 101 (Almost) Philosophical Positions

The first of these positions, of course, was the cybernetic realism of Chapters 2 and 3. The task of those two early chapters was to situate cybernetic realism in the context of materialist ontology and, through our analysis of the first person and third person problems, to see some of the rationale behind it. Without some such position—specifically, that there are matters of fact about what it is like to *be* certain material structures—the rest of this dissertation, as a project aimed at furthering the cause of a comprehensive science of the mind based on material monism, could never have got off the ground. (After all, to deny that there are matters of fact about what it is like to be certain material structures sounds either like the denial of material monism or the denial that there is anything which

it is like something to be, full stop!) To that extent, our position on the issue was entirely helpful.

Again in Chapter 4, we took and eventually defended against Brian Josephson a position the likes of which was indispensable to the task we've set ourselves of examining mental questions through explorations of the physical. If it should turn out to be *incorrect* that consciousness is irrelevant to quantum measurement, then the cognitive scientist has one of two choices. On the one hand, the scientist might "give up the ghost", so to speak, and allow that there is a non-material realm of mind which somehow interacts with the physical world and brings about state vector reduction. Under this option, many of the interesting questions about minds are spirited off into a realm where the empirical methods of physics are almost certain to fail us in our bid to understand the mental. The mind in that case would then be destined probably to remain forever as that last bastion of mental privacy we described in the Introduction. On the other hand, the cognitive scientist might instead "bite the bullet" and take on the burden of explaining just how it is that some material structures collapse wavepackets by virtue of being conscious while others do not (or, alternatively, how it is that conscious physical structures are prevented from existing in states of linear superposition while nonconscious ones are not). This would be a brave undertaking which I, for one, and which apparently no other author in this field, has any clear idea how to approach. Thus not only was our position on the matter helpful, but let's hope for the sake of everyone pursuing questions of material cognition that it's correct!

In the next chapter, we employed cybernetic realism, now hopefully safe from one possible criticism from quantum physics, as the backdrop for a new position on how consciousness actually emerges from the material substrate. The self model approach suggests that what Dennett calls the "center of narrative gravity", or what we term "the self", lies in the data structure which may be instantiated by particular material structures. I expressed the hope in that chapter that the self model view might help clarify the arguably problematic notion of self, and I wondered about how the sceptic might debate with someone who had thought of themselves for all their lives as a self model.

On the first issue, I believe the self model approach has helped clarify the notion of the self, and it has certainly offered new ways of

addressing some familiar philosophical issues surrounding the subjects of conscious experience. In particular, it is at the centre of a particularly useful treatment of qualia which suggests that qualities of experience are always relational and can be understood through the relationships internal to the materially instantiated self model data structure. That is, the self model view offers us somewhere to embody these qualities of experience; later we saw how it might also offer us some handles on the grain problem and frame problem, as well as a way of addressing Putnam's points about meaning and even a speculative way of accommodating neurophysiological evidence about oscillations in cortical activity.

In the course of exploring the self model idea in some detail, we noted a number of characteristics which we would expect self models to have, and we saw both how some of those characteristics might be implemented neurally as well as how some of those characteristics may help transcend some known properties of neural architectures (such as bounds on dendritic and axonal arborisation). The emergence of polymodal associations in self model data structures also gave some insight into the development of language, while several self model characteristics helped us on the way to overcoming some commonly sited shortcomings of functionalist views of mind.

In Chapter 8, we checked on the evolutionary plausibility of the self model view and noted some of the selective advantages which we would expect for organisms equipped with a self model. We returned to the language theme and speculated on how language emergent from a self model might lend itself to improved procedural memory, and we noted one possible explanation of the "internal dialogue". We also saw how self models may speed the learning process, and through a martial arts example we traced very broadly a possible sequence of changes in an organism equipped with a self model as it learns a particularly complex set of behaviours.

All told, the self model view has turned out to be a rich point of departure for investigating a host of philosophical and cognitive scientific questions, and we've demonstrated at least its *prima facie* biological plausibility, taking note of several more of its advantages along the way. We also saw, in Chapter 9, how it coheres with Edelman's low level theory of perceptual categorisation and helps to bridge the gap between it and higher level phenomenal experience. Since rounding out the self model

discussion with artificial neural network examples and comments on information processing, the self model view appears to suggest a robust set of approaches which, while certainly not yet as well supported as the existence of *w*-particles, nonetheless shows promise as a fruitful topic for further investigation. What to say to the hypothetical individual who thinks of himself or herself as a materially instantiated data structure, I leave to the sceptic to sort out.

In the second half of the dissertation, we took what amounts to the rather safe position that philosophically relevant things can be said about the dynamics of a low level material substrate and the conscious experience of a self model instantiated by it (and, in particular, that specifically chaotic dynamics in that substrate might be especially important). While not contradicting our earlier argument that the self model is blind to a wide variety of characteristics of its instantiating hardware or wetware, this angle suggests simply that relationships embodied by, and the temporal evolution of, self models is largely dependent on dynamical properties of the instantiating substrate. As an aide to exploring this idea, we introduced a special representational schema and immediately saw how it could be applied to existing theories and how it might even help us to analyse questions about free will, creativity, and memory. It gave us a way of phrasing observations about the relationship of chaotic low level dynamics to the temporal evolution of higher level features of cognitive systems; in the next chapter it helped us to defend against Peter Smith an argument for indeterminism in cognitive state transitions based on observations about underlying dynamics and particular ways of mapping sets of low level states of a material substrate to higher level mental states. Hopefully the usefulness of both the original position and the representational schema emerged clearly at this early stage.

In the Chapter 15, we entered a technical discussion about the computability of a class of chaotic systems, and we found that the implication from the computability of a set of governing equations to the computability of the behaviour of a corresponding real system is far from straightforward. Our analysis suggested that there *may* be aspects of behaviour in chaotic analogue substrates which are not Turing computable. We noted that if this is true, may carry important implications for the way we view human cognition and its relationship to

computation. If true, the conclusions of this chapter underscore the importance of multi-level analysis and suggest that human minds may be significantly different to anything which might be produced by the strong AI project.

With such important implications at stake, we turned in the next chapter to addressing a series of comments due to Peter Smith which threaten to reduce to irrelevance the presence of specifically chaotic dynamics in the material substrate of cognitive systems. In this and the two subsequent chapters, we defended the view of chaos to which we had so far appealed against criticisms from infinite intricacy, predictability, and complexity. Along the way, we clarified a number of points about how chaos should be treated, we explored a real system which appears to vindicate the position taken in Chapter 15 on the computability of chaotic analogue systems, and we offered a new measure of complexity designed to overcome the problems of previous measures as well as helping in the defence against Smith. In the end, we wound up with positions largely compatible in most cases but in some important instances strictly incompatible with Smith's. Over the course of these three chapters, we legitimated our original interest in chaotic dynamics and its possible relationship to higher level instantiated data structures.

In the last of the main chapters of this dissertation, we returned to the issue of representation which played such a central rôle in our understanding of self model data structures, applying our new grasp of complexity to demonstrate first that complexity is inextricably bound up with representation and second that the part played by chaotic dynamics in intelligent systems is emphatically an open question. It so far suggests only possibilities and cannot be prejudged either by Smith's well-targeted but arguably overzealous naysaying or by the overly optimistic assumptions frequently made, for instance, in the popular press, that chaos somehow is the magical key to human escape from determinism and all other "problems" of material monism. Although this last sequence of chapters of philosophy of science was rather a roundabout way of getting to our destination, we can now see that our original position that it is important to investigate relationships between low level dynamics and higher level data structures—with particular attention to specifically chaotic dynamics—has been altogether vindicated.

20.3 Telling Them Again—The Short Form

Overall, we've covered a very wide range of topics, and each of the positions we've taken on substantive issues have proven themselves worthwhile—which shouldn't be too surprising, of course, since the positions which turned out to be less fruitful were deliberately omitted from this dissertation! We've seen the value of analysing cognitive systems at different levels of description and from different points of view, we've seen that quantum mechanics is largely irrelevant to the materialist cognitive science project, and we've seen how the self model approach to seating conscious may help answer a variety of interesting questions. I can't vouch for anyone else, but I feel I *am* a materially instantiated data structure! We've also noted how chaos theory may be important to philosophy of mind and how it should at least be kept an open subject for further study, and our exploration of complexity in the context of chaos helped us also to reveal how complexity cannot be divorced from representation. Along the way, we've made some promising speculations on topics as diverse as "empirical logic", the illusion of qualia, the internal dialogue, content addressable memory, the grain and frame problems, the rôle of cortical oscillations, creativity, free will, and others. So what *haven't* we covered, and what haven't we covered well enough?

20.4 Shortcomings of a House on Stilts

Perhaps it would take another book as long as this one to begin to cover the shortcomings of our analyses. The overall picture on offer here is like a house built on stilts, meant to raise us up out of the wet and muck of various philosophical quandaries, that from our various positions we might spy through the windows of our house some dry land on which to build a better philosophical framework. But while we've done a fair bit of direct checking that our stilts are in order, more of our time has been spent looking out the windows and suggesting that if the views are so good, then the stilts must be steady. While this is hardly very different to the physicist and her w-particles, it is worthwhile to take note explicitly of some of the issues which could benefit from further analysis.

First, there is certainly much more to be said about relationships between self model data structures and their neural wetware. We've also

not explored sufficiently the relationships between various parts of the self model and between the self model and other data structures instantiated by the material substrate but not directly incorporated in the self model. Surely other nonconscious models, in addition to the self model, are instantiated by neural subsystems; how is their rôle in the system different to that of the self model? And what particular changes occur in the self model during disruptions to normal conscious experience such as anaesthesia or intoxication?

Along similar lines, we might also ask what other information theoretic characteristics are displayed by self models apart from those we've explored? How might these characteristics be implemented with neural networks, and how might the very primitive example networks we've discussed be improved? What, if any, low level characteristics of these networks are absolutely essential, and which could be replaced by simpler ones?

Other significant questions about low level characteristics remain. It's clear, for instance, that we've not concluded the analysis of recursion theory and the importance of analogue chaos. That our tentative conclusions on the subject are at odds with almost universally accepted received wisdom in computation circles is perhaps good reason to believe something has gone wrong in our approach, although the failure so far of any readers or conference audiences to point out exactly where we've gone wrong is some reason to believe the opposite.

Later in our analysis of chaotic dynamics, we spent a good deal of time examining issues of complexity and even offering a new complexity measure. However, none of those discussions got us any closer actually to *measuring* complexity; since none of the complexity measures we discussed were computable, they served only to help us in theorising about complexity in the abstract. Actually measuring complexity is a difficult and actively researched topic, and while our analysis has contributed little to that research, it could certainly benefit from whatever developments may yet emerge from it.

Similarly, it is a shortcoming that all our explorations of chaotic dynamics have focussed on dynamical systems actually *in* a chaotic, or supercritical, regime. In fact, it is almost certain that the real payoff from chaotic systems in terms of information density and so forth will come from studying systems at the critical *transition* between more

straightforward order and chaos. This area, too, is currently the subject of active research, and more interesting philosophical insights will undoubtedly come from this research than from our own investigations of systems already at supercritical stages.

Finally, we have played very fast with the debate surrounding topics like qualia, free will, and so forth. It was never our purpose to come to any great resolutions of these debates, but nonetheless it is clear that more work needs to be done before it can really be accepted that the self model approach has useful bearing on these problems. What discussion we have undertaken, however, has at least suggested that these are areas where something relevant might be said, where the view on offer may well have some important impact.

20.5 New Directions, New Positions

Some of the new directions for research suggested by the material we've covered here come not from shortcomings in our analyses but rather from deliberate omissions. For instance, we've said nothing about the ethical implications of the self model view. It is interesting, for instance, that attributions of value to persons may partially reduce on this account to attributions of value to particular *patterns* of activity instantiated by material structures. This idea might form a useful complement to the Psychological Principle of Sufficient Reason I've described elsewhere. (Mulhauser 1993e) It may also bear on issues of both intrinsic and instrumental value in environmental ethics and animal rights.

We might also wonder about the implications of this kind of materialist view of cognition for questions in aesthetics. To what extent, for instance, are human perceptions of beauty predicated on complementarity or contrast with materially instantiated patterns in their self models? This question might also prompt us to wonder more about the suggestion in Chapter 3 that human perceptions of logic may derive from patterns contingently instantiated by their self models.

Another suggestion, with which we closed Chapter 3, might also merit more attention. We noted that scientific explanations always at some level come to describing *how* mechanisms operate but not *why* they operate that way (instead of some other way). In seating philosophy of

mind squarely within the boundaries of science, interesting philosophical questions might be asked about the significance of this ultimate end to scientific explanation. Even curious questions of theology might evolve from the notion that our best scientific explanations of minds ultimately reduce to descriptions of how they in fact are but not accounts of why they are.

Not far from this, perhaps, is the idea on which we first touched in the first section of the Introduction that if we assume a monist material ontology, it is all the more amazing that we materially instantiated consciousnesses sense or perceive anything at all! To suppose a separate realm of mind or spirit is to lock up our awe safe from the grasp of objective empirical inquiry, but to integrate that realm of mind or spirit into one coherent materialist ontology is to take hold wholeheartedly, with all our empirical capacity to hear and see and feel, of the utter amazingness of those very capacities. To think of a "me" in a separate mental realm who wonders about such things is not incredible, but to think of a "me" in a strictly material realm leaves "me" dumbstruck!

20.6 Buying Philosophical Futures

Investigating issues like these with methods akin to those we've employed here is, I believe, the future of philosophy of mind. Some philosophers privately express the fear that the march of cognitive science is slowly appropriating the territory formerly allotted to philosophical inquiry; but if it is really doing this, I believe it is doing so only at the same time that new territories are being opened for philosophy.

At least some of this new territory is given to philosophers pursuing, as we have largely done here, what amounts to "theoretical cognitive science", or theoretical AI. It is a job for philosophers to synthesize plausible theoretical accounts from existing data and to offer theoretical direction to the research projects of cognitive science and artificial intelligence. It is a job for philosophers to integrate, as we have tried to do here, information from a wide range of philosophical and scientific fields. The Grand Unification Theory of Consciousness, if there is ever to be such a thing, can no more emerge from and be couched in the terms of just one philosophical or scientific specialty than the Grand Unification Theory of physics can come from just the study of w-particles

and the weak nuclear force. If philosophers do not take the initiative in setting out the requirements for what must be explained by such theories of consciousness and in verifying the consistency and plausibility of candidate theories or bits of theories, then, I submit, no one will.

I hope the bits of theory and analysis offered here may be subjected to this kind of assessment by other philosophers. And, as I expressed at the end of the Introduction, I hope some of these ideas may yet find more refinement in their understandings of the world than they have in my own.

Appendix

Permissions to include a number of reprinted papers have been granted as below, with the text in quotation marks as suggested by the respective publishers. This is followed by a statement on the authorship of this dissertation and a note of its length in words.

- Mulhauser 1993a (from Chapter 2)

  "First published in *The Philosopher*, April 1993, pp. 19-24."

- Mulhauser 1993b (from Chapter 10)

  "Published in:
  Proceedings of the European Symposium on Artificial Neural Networks, Brussels (Belgium), April 1993, M. Verleysen ed.
  Publisher: D facto publications
  45 rue Masui
  B-1210 Brussels (Belgium)"

- Mulhauser 1993d and forthcoming (from Chapters 13 and 15)

  I have had severe difficulty ascertaining the status of this material but include it here with the understanding that I retain copyright because I have not entered into any agreement transferring copyright to another party. While I have received assurances from Prof. Nikolai Kossovsky of the State University of St. Petersburg that this piece will be published in the proceedings volume of the International Congress on Computer Systems and Applied Mathematics 1993, and Dr. Sergei Voitenko of the Centre for Modern Communications, the original congress organiser, has indicated the publisher will be Springer-Verlag, I have not been able to contact Prof. Kossovsky or Dr. Voitenko for more than half a year. Poor conditions in Russia have compromised even my communication with a close personal friend in St. Petersburg, and my attempts to contact Prof. Kossovsky or Dr. Voitenko have been entirely fruitless.

- Mulhauser 1994b (from Chapter 10)

"Published in:
Proceedings of the European Symposium on Artificial Neural Networks, Brussels (Belgium), April 1994, M. Verleysen ed.
Publisher: D facto publications
45 rue Masui
B-1210 Brussels (Belgium)"

- Mulhauser 1995 in press (from Chapter 4)

Forthcoming in *Minds and Machines*, February 1995.

I, Gregory R. Mulhauser, certify that with the exception of certain short quotations, appropriately cited, all text and illustrations in this dissertation, *Mind Out of Matter: Topics in the Physical Foundations of Consciousness and Cognition*, are entirely of my own creation. I have personally undertaken all the alterations which have been prompted by the normal course of reviewing and commenting by colleagues, many of whom are noted in the section titled Foreword and Acknowledgements.

Gregory R. Mulhauser 13 Sept. 1994

Signed *Date*

Word count for main text and footnotes
Ch. 1–20 (excluding bibliography, index, etc.): 99,794

Word count for entire document: 111,079

References

Aarts, E. and J. Korst. (1989) *Simulated annealing and Boltzmann machines.* New York: John Wiley and Sons.

Aberth, O. (1971) 'The failure in computable analysis of a classical existence theorem for differential equations', *Proceedings of the American mathematical society* 30: 151-156.

Ahlers, G. and R.W. Walden. (1980) 'Turbulence near onset of convection', *Physical review letters* 44: 445+.

Albert, David Z. (1983) 'On quantum-mechanical automata', *Physics letters* 98A: 249-252.

Albert, David Z. (1987) 'A quantum-mechanical automaton', *Philosophy of science* 54: 577-585.

Albert, David Z. (1990) 'The quantum mechanics of self-measurement', in *Complexity, entropy, and the physics of information*, SFI Studies in the Sciences of Complexity VIII, ed. Wojciech H. Zurek, pp. 471-476. Redwood City, California: Addison-Wesley.

Albrecht, Andreas. (1992a) 'Following a "collapsing" wavefunction', *Fermilab report* (unpublished).

Albrecht, Andreas. (1992b) 'Investigating decoherence in a simple system', *Physical review* D46: 5504-5520.

Alexander, J.C.; J.A. Yorke, Z. You, and I. Kan. (1992) *International journal of bifurcation and chaos* 2: 795-813.

Ambros-Ingerson, J.; R. Granger, and G. Lynch. (1990) 'Simulation of paleocortex performs hierarchical clustering', *Science* 247: 1344-1348.

Arensburg, Baruch; L.A. Schepartz, Anne-Marie Tillier, Bernard Vandermeersch, and Yoel Rak. (1990) 'A reappraisal of the anatomical basis for speech in middle paleolithic hominids', *American journal of physical anthropology* 80: 137-146.

Banks, J.; J. Brooks, G. Cairns, G. Davis, and P. Stacey. (1992) 'On Devaney's definition of chaos', *American mathematical monthly*, April: 332-334.

Barahona, F. (1982) 'On the computational complexity of Ising spin glass models', *Journal of physics* A15: 3241-3253.

Barlow, H.B. (1980) 'Nature's joke: A conjecture on the biological role of consciousness', in *Consciousness and the physical world*, ed. B.D. Josephson and V.S. Ramachandran. Oxford: Pergamon press.

Barnsley, Michael. (1988) *Fractals everywhere.* San Diego: Academic Press.

Bekenstein, J.D. (1981) *Physical review* D23: 287+.

Benioff, Paul. (1980) *International journal of statistical physics* 22: 563+.

Benioff, Paul. (1982) *International journal of statistical physics* 29: 515+.

Bennett, C.H. (1973) *IBM journal of research and development* 17: 525+.

Bennett, C.H. (1982) 'The thermodynamics of computation—a review', *International journal of theoretical physics* 21: 905-940.

Bennett, C.H. (1987) 'Information, dissipation, and the definition of organization', in *Emerging syntheses in science*, ed. David Pines. Reading, Massachusetts: Addison-Wesley.

Bennett, C.H. (1990) 'How to define complexity in physics, and why', in *Complexity, entropy, and the physics of information*, SFI studies in the sciences of complexity

VIII, ed. Wojciech H. Zurek, pp. 137-148. Redwood City, California: Addison-Wesley.
Bergé, Pierre; Yves Pomeau, and Christian Vidal. (1984) *Order within chaos.* New York: John Wiley and Sons.
Blackmore, Susan. (1993) *Dying to live.* London: Grafton.
Block, N. (1980) 'Are absent qualia impossible?', *Philosophical review* 89: 257-274.
Boden, Margaret. (1990) *Philosophy of AI.* Oxford: Oxford University Press.
Bohm, David. (1952) 'A suggested interpretation of the quantum theory in terms of "hidden" variables', *Physical review* 85: 166-193.
Broadbent, D. (1985) 'A question of levels: Comment on McClelland and Rumelhart', *Journal of experimental psychology (general)* 114: 189-192.
Brodal, A. (1981) *Neurological anatomy in relation to clinical medicine.* Oxford: Oxford University Press.
Browne, Antony and John Pilkington. (1994) 'Variable binding in a neural network using a distributed representation', in *European symposium on artificial neural networks 1994*, ed. Michel Verleysen, pp. 199-204. Brussels: D facto.
Burge, T. (1979) 'Individualism and the mental', *Midwest studies in philosophy* 4: 73-122.
Burge, T. (1982) 'Other bodies', in *Thought and object*, ed. A. Woodfield. Oxford: Oxford University Press.
Burge, T. (1986) 'Intellectual norms and foundations of mind', *Journal of philosophy* 83: 697-720.
Carleton, L. (1983) 'The population of China as one mind', *Philosophy research archives* 9: 665-674.
Chaitin, G.J. (1994) 'Randomness & complexity in pure mathematics', *International journal of bifurcation and chaos* 4(1). (Retrievable from International Philosophical Preprint Exchange, phil-preprints.l.chiba-u.ac.jp.)
Chalmers, D.J. (1990) 'Syntactic transformations on distributed representations', *Connection science* 2(1 & 2): 53-62.
Changeux, J.-P. and A. Danchin (1976) 'Selective stabilization of developing synapses as a mechanism for the specification of neuronal networks', *Nature* 264: 705-711.
Changeux, J.-P.; T. Heidmann, and P. Patte (1984) 'Learning by selection', in *The biology of learning*, ed. P. Marler and H.S. Terrace, pp. 115-137. New York: Springer-Verlag.
Chapeau-Blondeau, F. (1993) 'Analysis of neural networks with chaotic dynamics', *Chaos solitons & fractals* 3: 133-139.
Choi, M.Y. and B.A. Huberman. (1983) 'Dynamic behavior of nonlinear networks', *Physical review* A28: 1204-1206.
Chrisman, L. (1991) 'Learning recursive distributed representation for holistic computation', *Connection science* 3(4): 345-365.
Churchland, Patricia S. (1986) *Neurophilosophy: Toward a unified science of the mind-brain.* Cambridge, Massachusetts: MIT Press.
Churchland, Patricia S. and Terrence J. Sejnowski. (1992) *The computational brain.* Cambridge, Massachusetts: MIT Press.
Clark, A. (1989) *Microcognition.* Cambridge, Massachusetts: MIT Press.
Clark, A. and R. Lutz, eds. (1992) *Connectionism in context.* New York: Springer-Verlag.
Clement, B.E.P.; P.V. Coveney, M. Jessel, and P.J. Marcer. (1991) 'The brain as a Huygens machine', in *Nature, cognition and system*, vol. 3, ed. M. Carvallo. Dordrecht: Kluwer Academic Press.
Coleman, S.; J. Hartle, T. Piran, and S. Wienberg, eds. (1991) *Quantum cosmology and baby universes: Proceedings of the 7th Jerusalem Winter School*, Jerusalem, Israel, 1990. Singapore: World Scientific.
Corbi, J.E. (1993) 'Classical and connectionist models: Levels of description', *Synthese* 95: 141-168.
Crick, Francis. (1984) 'Function of the thalamic reticular complex: the searchlight hypothesis', *Proceedings of the national academy of sciences U.S.A.* 81: 4586-4590.
Crick, Francis. (1994) *The astonishing hypothesis: The scientific search for the soul.* New York: Simon & Schuster.

Crick, Francis and Christof Koch. (1990) 'Toward a neurobiological theory of consciousness', *Seminars in the neurosciences* 2: 263-275.
Crick, Francis and Christof Koch. (1992) 'The problem of consciousness', *Scientific American* 267: 111-117.
Cuda, T. (1985) 'Against neural chauvinism', *Philosophical studies* 48: 111-127.
Cytowic, Richard E. (1989) *Synesthesia: A union of the senses.* Springer series in neuropsychology. New York: Springer-Verlag.
Cytowic, Richard E. (1993) *The man who tasted shapes: A bizarre medical mystery offers revolutionary insights into emotions, reasoning & consciousness.* New York: G.P. Putnam's.
Darwin, C. (1859) *On the origin of species by means of natural selection or the preservation of favoured races in the struggle for life.* London: Murray.
Darwin, C. (1872) *The expression of emotions in man and animals.* London: Murray.
Davies, P.C.Q. (1981) 'Is thermodynamic gravity a route to quantum gravity?', in *Quantum gravity 2: A second Oxford symposium*, ed. C.J. Isham, R. Penrose and D.W. Sciama, pp. 183-209. Oxford: Clarendon Press.
Davis, L. (1982a) 'What is it like to be an agent?', *Erkenntnis* 18: 195-213.
Davis, L. (1982b) 'Functionalism and absent qualia', *Philosophical studies* 41: 231-249.
Dennett, Daniel C. (1984) *Elbow room: The varieties of free will worth wanting.* Oxford: Clarendon Press.
Dennett, Daniel C. (1987) *The intentional stance.* London: MIT Press.
Dennett, Daniel C. (1988) 'Quining qualia', in *Consciousness in contemporary science*, ed. A.J. Marcel and E. Bisiach, pp. 42-77. Oxford: Clarendon Press.
Dennett, Daniel C. (1991) *Consciousness explained.* Boston: Little, Brown.
Der, R. and Herrmann, M. (1994) 'Instabilities in self-organized feature maps with short neighbourhood range', in *European symposium on artificial neural networks 1994*, ed. Michel Verleysen, pp. 271-276. Brussels: D facto.
Desmond, N.L. and W.B. Levy (1981) 'Ultrastructural and numerical alteration in dendritic spines as a consequence of long-term potentiation', *Anatomical record*, 199: 68A-69A
Deutsch, D. (1985a) 'Quantum theory as a universal physical theory', *International journal of theoretical physics* 24: 1-41.
Deutsch, D. (1985b) 'Quantum theory, the Church-Turing principle and the universal quantum computer', *Proceedings of the royal society of London* A400: 97-117.
Deutsch, D. (1989) *Proceedings of the royal society of London* A425: 73+.
Devaney, Robert L. (1988) *An introduction to chaotic dynamical systems.* New York: Addison-Wesley.
Duchateau, G. and Anders Lansner. (1991) 'A Bayesian artificial neural network with spiking units', *Report TRITA-NA-P9101*, Royal Institute of Technology (Sweden).
Eccles, J. (1986) 'Do mental events cause neural events analogously to the probability fields of quantum mechanics?', *Proceedings of the royal society of London* B227: 411-428.
Eccles, J. (1990) 'A unitary hypothesis of mind-brain interaction in the cerebral cortex', *Proceedings of the royal society of London* B240: 433-451.
Edelman, Gerald M. (1978) 'Group selection and phasic reentrant signaling: A theory of higher brain function', in *The mindful brain: Cortical organization and the group-selective theory of higher brain function*, by G.M. Edelman and V.B. Mountcastle, pp. 51-100. Cambridge, Massachusetts: MIT Press.
Edelman, Gerald M. (1981) 'Group selection as the basis for higher brain function', in *Organization of the cerebral cortex*, ed. F.O. Schmitt, F.G. Worden, G. Adelman, and S.G. Dennis, pp. 535-563. Cambridge, Massachusetts: MIT Press.
Edelman, Gerald M. (1989a) *Neural Darwinism: The theory of neuronal group selection.* Oxford: Oxford University Press.
Edelman, Gerald M. (1989b) *The remembered present: A biological theory of consciousness.* New York: Basic Books.

Edelman, Gerald M. and G.N. Reeke, Jr. (1982) 'Selective networks capable of representative transformation, limited generalizations, and associative memory', *Proceedings of the national academy of sciences U.S.A.* 79: 2091-2095

Ede'man, Gerald M. and L.H. Finkel (1984) 'Neuronal group selection in the cerbral cortex', in *Dynamic aspects of neocortical function*, ed. G.M. Edelman, W.E. Gall, and W.M. Cowan, pp. 653-695. New York: Wiley.

Edelman, Gerald M.; W.E. Gall, and W.M. Cowan, eds. (1985) *Molecular bases of neural development.* New York: Wiley.

Eeckman, F.H. and W.J. Freeman. (1991) 'Asymmetric sigmoid nonlinearity in the rat olfactory system', *Brain research* 557: 13-21.

Everett, H. (1957) '"Relative state" formulation of quantum mechanics', *Reviews of modern physics* 29: 454-462.

Fan, Y.S. and A.V. Holden. (1993) 'Bifurcations, burstings, chaos and crises in the Rose-Hindmarsh model for neuronal activity', *Chaos solitons & fractals* 3: 439-449.

Fatmi, H.E. and G. Resconi. (1988) 'A new computing principle', *Il nuovo cimento* 101b: 239-242.

Fatmi, H.E.; M. Jessel, P.J. Marcer, and G. Resconi. (1990) 'Theory of cybernetic and intelligent machines based on Lie commutators', *International journal of general systems* 16: 123-164.

Feynman, R. (1986) 'Quantum mechanical computers', *Foundations of physics* 16: 507-531.

Fifková, E. and A. van Harreveld (1977) 'Long-lasting morphological changes in dendritic spines of dentate granular cells following stimulation of the entorhinal area', *Journal of neurocytology* 6: 211-230

Finkel, L.H. and G.M. Edelman (1985) 'Interaction of synaptic modification rules within populations of neurons', *Proceedings of the national academy of sciences U.S.A.* 82: 1291-1295

Flanagan, O.J. (1985) 'Consciousness, naturalism and Nagel', *Journal of mind and behavior* 6: 373-390.

Fodor, J.A. and B.P. McLaughlin. (1990) 'Connectionism and the problem of systematicity: Why Smolensky's solution did not work', *Cognition* 35: 183-204.

Fodor, J.A. and Z. Pylyshyn. (1988) 'Connectionism and cognitive architecture: A critical analysis', *Cognition* 28: 3-71.

Ford, K.M. and P.J. Hayes. (1991) *Reasoning agents in a dynamic world: The frame problem.* Greenwich: JAI Press.

Foss, J. (1989) 'On the logic of what it is like to be a conscious subject', *Australasian journal of philosophy* 67: 305-320.

Foster, Sara and Harvey R. Brown. (1988) 'On a recent attempt to define the interpretation basis in the many worlds interpretation of quantum mechanics', *International journal of theoretical physics* 27: 1507-1531.

Fredkin, E. and E. Toffoli. (1982) *International journal of theoretical physics* 21: 219+.

Freeman, W. J. (1989) 'Analysis of strange attractors in EEGs with kinesthetic experience and 4-D computer graphics', in *Brain dynamics*, ed. E. Basar and T.H. Bullock. Heidelberg: Springer.

Freeman, W.J. (1964) 'A linear distributed feedback model for prepyriform cortex', *Experimental neurology* 10: 525-547.

Freeman, W.J. (1972) 'Measurement of open-loop responses to electrical stimulation in olfactory bulb of cat', *Journal of neurophysiology* 35: 745-761.

Freeman, W.J. (1975) *Mass action in the nervous system.* New York: Academic Press.

Freeman, W.J. (1979) 'Nonlinear gain mediation of cortical stimulus response relations', *Biological cybernetics* 33: 237-247.

Freeman, W.J. (1987a) 'Techniques used in the search for the physiological basis of the EEG', in *Handbook of electroencephalography and clinical neurophysiology*, vol. 3A, eds. A.S. Gevins and A. Remond. Amsterdam: Elsevier.

Freeman, W.J. (1987b) 'Simulation of chaotic EEG patterns with dynamic model of the olfactory system', *Biological cybernetics* 56: 139-150.

Freeman, W.J. (1988) 'Strange attractors that govern mammalian brain dynamics shown by trajectories of electroencephalographic (EEG) potential', *IEEE transactions on circuits and systems* 35: 781-783.

Freeman, W.J. (1991a) 'Predictions on neocortical dynamics derived from studies in paleocortex', in *Induced rhythms of the brain*, eds. E. Basar and T.H. Bullock. Cambridge, Massachusetts: Birkhaeuser Boston.

Freeman, W.J. (1991b) 'The physiology of perception', *Scientific American* 264: 78-85.

Freeman, W.J. and C.A. Skarda. (1985) 'Spatial EEG patterns, nonlinear dynamics and perception: the neo-Sherringtonian view', *Brain research reviews* 10: 147-175.

Gabor, D.; W.P.L. Wilby, and R. Woodcock. (1960) 'A universal non-linear filter, predictor and simulator, that optimizes itself by a learning process', *Proceedings of the IEEE* 108B: 422-438.

Gardner, Daniel, ed. (1993) *The neurobiology of neural networks.* Cambridge, Massachusetts: MIT Press.

Garey, M.R. and D.S. Johnson. (1979) Computers and intractability: A guide to the theory of NP-completeness. San Francisco, California: Freeman.

Gell-Mann, M. and J.B. Hartle. (1990) 'Quantum mechanics in the light of quantum cosmology', in *Complexity, entropy, and the physics of information*, SFI Studies in the Sciences of Complexity VIII, ed. Wojciech H. Zurek, pp. 425-469. Redwood City, CA: Addison-Wesley.

Geschwind, Norman. (1964) 'The development of the brain and the evolution of language', in *Monograph series on language and linguistics*, Vol. 17, ed. C.I.J.M. Stuart. Georgetown: Georgetown University Press.

Geschwind, Norman. (1965) 'Disconnexion syndromes in animals and man', *Brain* 88: 237-294+.

Gluck, Mark A.; D.B. Partker, and E.S. Reifsnider. (1989) 'Learning with temporal derivatives in pulse-coded neuronal systems', *Neural information processing systems* 1: 195-203.

Gluck, Mark A. and David E. Rumelhart, eds. (1990) *Neuroscience and connectionist theory.* London: Lawrence Erlbaum Associates, Inc.

Goertzel, B. (1993) *The evolving mind.* New York: Gordon and Breach.

Graham, A.C. (1981) *Chuang-tzu: The seven inner chapters.* London: George Allen & Unwin.

Gray, C.M.; P. Konig, A.K. Engel, and W. Singer. (1989) 'Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties', *Nature* 338: 334-337.

Grebogi, C.; E. Kostelich, E. Ott, and J.A. Yorke. (1987) 'Multi-dimensioned intertwined basin boundaries—basin structure of the kicked double rotor', *Physica* D25: 347-360.

Greenfield, Patricia M. (1992) 'Language, tools, and brain: The ontogeny and phylogeny of hierarchically organized sequential behavior', *Behavioral and brain sciences* 14: 531-595.

Grzegorczyk, A. (1955) 'Computable functionals', *Fundamentals of mathematics* 42: 168-202.

Grzegorczyk, A. (1957) 'On the definitions of computable real continuous functions', *Fundamentals of mathematics* 44: 61-71.

Gustafsson, Mats; Lars Asplund, Olle Gällmo, and Ernst Nordström. (1992) 'Pulse coded neural networks for hardware implementation', Presented 9 September 1992 at the *First Swedish national conference on connectionism* in Skövde, Sweden.

Haksar, V. (1981) 'Nagel on subjective and objective', *Inquiry* 24: 105-121.

Harnad, S. (1987) 'The induction and representation of categories', in *Categorical perception: The groundwork of cognition*, ed. S. Harnad. New York: Cambridge University Press.

Harnad, S. (1990) 'The symbol grounding problem', *Physica* D 42: 335-346.

Harnad, S. (1992) 'Connecting object to symbol in modeling cognition', in *Connectionism in context*, ed. A. Clarke and R. Lutz. New York: Springer Verlag.

Harnad, S. (1993) 'Problems, problems: The frame problem as a symptom of the symbol grounding problem', *PSYCOLOQUY* 4(34) frame-problem.11.harnad.

Harnad, S.; S.J. Hanson, and J. Lubin. (1991) 'Categorical perception and the evolution of supervised learning in neural nets', in *Working papers of the AAAI spring symposium on machine learning of natural language and ontology*, ed. D.W. Powers and L. Reeker, pp. 65-74. Presented March 1991 at the *Symposium on symbol grounding: Problems and practice* at Stanford University.

Hayek, F.A. (1952) *The sensory order: An inquiry into the foundations of theoretical psychology.* Chicago: University of Chicago Press. (Midway reprint, 1976)

Hebb, D.O. (1949) *The organization of behavior: A neuropsychological theory.* New York: Wiley.

Hebb, D.O. (1980) *Essay on mind.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Hebb, D.O. (1982) 'Elaborations on Hebb cell assembly theory', in *Neuropsychology after Lashley*, ed. J. Orback, pp. 483-496. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Hille, B. (1984) *Ionic channels of excitable membranes.* Sunderland, Massachussetts: Sinauer Associates.

Hobbs, Jesse. (1991) 'Chaos and indeterminism', *Canadian journal of philosophy* 21: 141-164.

Hobbs, Jesse. (1994) 'Chaos and computational theories of mind', unpublished manuscript.

Hofstadter, Douglas R. (1981) 'Reflections on Nagel's What is it like to be a bat?', in *The mind's I*, ed. Douglas R. Hofstadter and Daniel C. Dennett, pp. 403-414. Sussex: Harvester Press.

Hopfield, J.J. and D.W. Tank. (1985) '"Neural" computation of decisions in optimization problems', *Biological cybernetics* 52: 1-12.

Horgan, Terence and John Tienson, eds. (1991) *Connectionism and the philosophy of mind.* Dordrecht: Kluwer Academic Press.

Horgan, Terence and John Tienson. (1993) 'Levels of description in nonclassical cognitive science', *PHILOSOPHY* 34, Royal Institute of Philosophy Supplement, pp. 159-188. Also forthcoming in *Philosophy and cognitive science*, ed. Christopher Hookway and Donald Peterson. Cambridge: Cambridge University Press. Note that this is the published version of 'Levels of description in connectionism: cognitive transition, dynamical system, network implementation', Presented 12 September 1992 at the *Royal institute of philosophy: Philosophy and the cognitive sciences* conference in Birmingham, England.

Horgan, Terence and John Tienson. (in press) 'A nonclassical framework for cognitive science', *Synthese:* special issue on connectionism and philosophy of mind, ed. A. Clark.

Hubel, D.H. and T.N. Wiesel (1977) 'Functional architecture of macaque monkey visual cortex', *Proceedings of the royal society of London* B198: 1-59.

Humphrey, Keith; Richard C. Tees, and Janet Werker. (1979) 'Auditory-visual integration of temporal relations in infants', *Canadian journal of psychology* 33: 347-352.

Humphrey, Nicholas. (1984) *Consciousness regained.* Oxford: Oxford University Press.

Hunt, G.M.K. (1987) 'Determinism, predictability and chaos', *Analysis* 47: 129-133.

Ingvar, D.H. (1985) '"Memory of the future': An essay on the temporal organization of conscious awareness", *Human neurobiology* 4: 127-136.

Jackendoff, Ray. (1983) *Semantics and cognition.* Cambridge, Massachusetts: MIT Press.

Jackendoff, Ray. (1987) 'The status of thematic relations in linguistic theory', *Linguistic Inquiry* 18: 369-411.

Jackendoff, Ray. (1990) *Semantic structures.* Cambridge, Massachusetts: MIT Press.

Jackson, F. (1982) 'Epiphenomenal qualia', *Philosophical quarterly* 32: 127-136.

Johnson, D.S. (1983) 'The NP-completeness column: An ongoing guide' *Journal of algorithms* 4: 87-100.

Josephson, Brian D. (1993a) Private email communication, 10 November.

Josephson, Brian D. (1993b) Private email communication, 10 November.

Josephson, Brian D. (1993c) Private email communication, 11 November.

Josephson, Brian D. (1993d) Private email communication, 25 November.
Kaneko, K. (1990) 'Globally coupled chaos violates the law of large numbers but not the central limit theorem', *Physical review letters* 65: 1391-1394.
Kekes, J. (1977) 'Physicalism and subjectivity', *Philosophy and phenomenological research* 37: 533-536.
Kellert, Stephen. (1993) *In the wake of chaos.* Chicago: University of Chicago Press.
Kiefer, Claus. (1991) 'Interpretation of the decoherence functional in quantum cosmology', *Classical and quantum gravity* 8: 379-391.
King, C.C. (1991) 'Fractal and chaotic dynamics in nervous systems', *Progress in neurobiology* 36: 279-308.
Kirkpatrick, S.; C.D. Gelatt, Jr.; and M.P. Vecchi. (1983) 'Optimization by simulated annealing', Science 220: 671-680.
Kleene, S.C. (1952) *Introduction to metamathematics.* Amsterdam: North Holland.
Kohonen, T. (1984) *Self-organization and associative memory.* New York: Springer-Verlag.
Kolen, J.F. and J.B. Pollack. (1990) 'Back propagation is sensitive to initial conditions', *Complex systems* 4,3: 269-280.
Kong, S.G. and B. Kosko. (1991) 'Differential competitive learning for centroid estimation and phoneme recongition', *IEEE Transactions on neural networks* 2: 118-124.
Kosko, B. (1992) *Neural networks and fuzzy systems: A dynamical approach to machine intelligence.* Englewood Cliffs, New Jersey: Prentice-Hall International.
Kripke, S. (1971) 'Naming and necessity', in *Semantics of natural language*, ed. D. Davidson and G. Harman, pp. 253-355+. Dordrecht: Reidel.
Kuffler, S.W.; J.G. Nicolls, and A.R. Martin. (1984) *From neuron to brain.* Sunderland, Massachussetts: Sinauer Associates.
Lacombe, D. (1955a) 'Extension de la notion de fonction récursive aux fonctions d'une ou plusieurs variables réelles I', *C.R. Académie de science, Paris* 240: 2478-2480.
Lacombe, D. (1955b) 'Extension de la notion de fonction récursive aux fonctions d'une ou plusieurs variables réelles II, III', *C.R. Académie de science, Paris* 241: 13-14+.
Landauer, Rolf. (1991) 'Information is physical', *Physics today* 44: 23-29.
Langton, C.G. (1992) 'Artificial life', in *1991 Lectures in complex systems*, ed. L. Nadel and D. Stein. Reading, Massachusetts: Addison-Wesley.
Leff, H.S. and A.F. Rex (1990) *Maxwell's demon: Entropy, information, computing.* Princeton: Princeton University Press.
Li, Z. and J.J. Hopfield. (1989) 'Modeling the olfactory bulb and its neural oscillatory processings', *Biological cybernetics* 61: 379-392.
Lieberman, Philip. (1984) *The biology and evolution of language.* Cambridge, Massachusetts: Harvard University Press.
Lieberman, Philip. (1985). 'On the evolution of human syntactic ability: Its pre-adaptive bases—motor control and speech', *Journal of human evolution* 14: 657-668.
Lieberman, Philip. (1989) 'Some biological constraints on universal grammar and learnability', in *The teachability of language*, ed. Mabel L. Rice and Richard L. Schiefelbusch. Edinburgh: Edinburgh University Press.
Liljenström, H. and M. Hasselmo. (1992) 'Acetylcholine and cortical oscillatory activity', in *Proceedings of the first annual conference on computation and neural systems*, San Francisco.
Lockwood, Michael. (1990) *Mind, brain, and the quantum.* Oxford: Basil Blackwell. (First published in 1989.)
Mackay, D.M. (1971) 'Scientific beliefs about oneself', in *The proper study*, Royal Institute of Philosophy Lectures 4, ed. G.N.A. Vesey, pp. 48-63. London: Macmillan.
Mackay, D.M. (1980) 'Conscious agency with unsplit and split brains', in *Consciousness and the physical world*, ed. B.D. Josephson and V.S. Ramachandran, pp. 95-113. Oxford: Pergamon Press.
Mahowald, Misha and Rodney Douglas. (1991) 'A silicon neuron', *Nature* 354.
Malcolm, N. (1988) 'Subjectivity', *Philosophy* 63: 147-160.

Marcer, Peter J. (1992) 'The conscious machine and the quantum revolution in information technology', *Kybernetes* 21(1): 18-22.
Margolus, Norman. (1986) 'Quantum computation', *Annals of the New York academy of science* 480: 487-497.
Margolus, Norman. (1990) 'Parallel quantum computation', in *Complexity, entropy, and the physics of information*, SFI Studies in the Sciences of Complexity VIII, ed. Wojciech H. Zurek, pp. 273-287. Redwood City, CA: Addison-Wesley.
Marks, Lawrence E.; Robin J. Hammeal, Marc H. Bornstein, Linda B. Smith. (1987) *Perceiving similarity and comprehending metaphor.* Monographs of the society for research in child development, vol. 52. Chicago: University of Chicago Press.
Marr, David. (1982) *Vision.* New York: Freeman.
Maxwell, Nicholas. (1988) 'Quantum propensition theory: A testable resolution of the wave/particle dilemma', *British journal for the philosophy of science* 39: 1-50.
Mayr, E. (1982) *The growth of biological thought: Diversity, evolution, and inheritance.* Cambridge, Massachusetts: Harvard University press.
McClamrock, R. (1992) 'Irreducibility and subjectivity', *Philosophical studies.*
McClelland, J.L. and D.E. Rumelhart. (1985) 'Levels indeed! A response to Broadbent', *Journal of experimental psychology (general)* 114: 193-197.
McCormick, David A. (1990) 'Membrane properties and neurotransmitter actions', in *The synaptic organization of the brain*, 3rd ed., ed. Gordon M. Sheperd, pp. 32-66. Oxford: Oxford University Press.
McCulloch, G. (1988) 'What it is like', *Philosophical quarterly* 38: 1-19.
McKenna, Terence K. and Timothy Ely. (1992) *Synesthesia.* New York City: Granary Books.
McLaughlin, B.P. (1991) 'The connectionism/classicism battle to win souls', in *Philosophy and the cognitive sciences*, ed. Christopher Hookway and Donald Peterson. Cambridge: Cambridge University Press. Note that this is the published version of the paper presented September 1992 at the *Royal institute of philosophy: Philosophy and the cognitive sciences* conference in Birmingham, England.
Merzenich, M.M.; J.H. Kaas; J.T. Wall, R.J. Nelson, M. Sur, and D.J. Felleman (1983a) 'Topographic reorganization of somatosensory cortical areas 3b and 1 in adult monkeys following restricted deafferentation', *Neuroscience* 8: 33-55.
Merzenich, M.M.; J.H. Kaas, J.T. Wall, R.J. Nelson, M. Sur, and D.J. Felleman (1983b) 'Progression of change following median nerve section in the cortical representation of the hand in areas 3b and 1 in adult owl and squirrel monkeys', *Neuroscience* 10: 639-665.
Merzenich, M.M.; R.J. Nelson, M.P. Stryker, M. Cynader, A. Schoppman, and J.M. Zook (1984) 'Somatosensory cortical map changes following digit amputation in adult monkeys', *Journal of comparative neurology* 224: 591-605.
Metzinger, Thomas. (1993) 'Subjectivity and mental representation', presented 4 July 1993 at the *Second annual conference of the European society for philosophy and psychology* in Sheffield, England.
Minsky, M. (1967). *Computation: finite and infinite machines.* Englewood Cliffs, New Jersey: Prentice-Hall.
Minsky, M. and S. Papert (1969) *Perceptrons: An introduction to computational geometry.* Cambridge: MIT Press.
Mountcastle, V.B. (1978) 'An organizing principle for cerebral function: The unit module and the distributed system', in *The mindful brain: Cortical organization and the group-selective theory of higher brain function*, by G.M. Edelman and V.B. Mountcastle, pp. 7-50. Cambridge, Massachusetts: MIT Press.
Mulhauser, Gregory R. (1991) 'Responsibility and freedom in a chaotic world', *The Willamette dialogue* 1: 1-10.
Mulhauser, Gregory R. (1992) 'Computability in neural networks' (early version), presented September 1992 at the meeting of the *British society for the philosophy of science* in Durham, England.

Mulhauser, Gregory R. (1993a) 'What is it like to be Nagel?', *The philosopher: Journal of the philosophical society of England* April: 19-24. Note that this is a revised version of the paper of the same name presented at the December 1992 meeting of the *Scottish philosophical forum* (now the Scottish Postgraduate Philosophical Association) in Edinburgh, Scotland.

Mulhauser, Gregory R. (1993b) 'Population coding in a theoretical biologically plausible network', in *European symposium on artificial neural networks 1993*, ed. Michel Verleysen, pp. 65-70. Brussels: D facto.

Mulhauser, Gregory R. (1993c) 'Chaotic dynamics and introspectively transparent brain processes', Presented 4 July 1993 at the *Second annual conference of the European society for philosophy and psychology* in Sheffield, England.

Mulhauser, Gregory R. (1993d and forthcoming) 'Computability in chaotic analogue systems', Presented 21 July 1993 at the *International congress on computer systems and applied mathematics* in St. Petersburg, Russia. Forthcoming in *CSAM 93*, Springer-Verlag.

Mulhauser, Gregory R. (1993e) 'A letter from a future self', Prepared and accepted for but withdrawn from 1993 *World congress on universalism* in Warsaw, Poland.

Mulhauser, Gregory R. (1993f) 'Cognitive transitions and the strange attractor: A reply to Peter Smith', Presented 7 September 1993 at the *Fifth joint council initiative summer school in cognitive science and human computer interaction* in Edinburgh, Scotland.

Mulhauser, Gregory R. (1994a) 'Computability in neural networks' (final version), Prepared and accepted for but withdrawn from 1994 *International mathematics and computers in simulation symposium on mathematical modelling* in Vienna, Austria.

Mulhauser, Gregory R. (1994b) 'Biologically plausible hybrid network design and motor control', in *European symposium on artificial neural networks 1994*, ed. Michel Verleysen, pp. 79-84. Brussels: D facto.

Mulhauser, Gregory R. (1995 in press) 'Materialism and the "problem" of quantum measurement', *Minds and machines*, forthcoming Feb. 1995 (accepted August 1993).

Nadel, Lynn; Lynn A. Cooper, Peter Culicover, and R. Michael Harnish, eds. (1989) *Neural connections, mental computation*. London: MIT Press.

Nagel, Thomas. (1979) 'What is it like to be a bat?', reprinted in *Mortal questions*, pp. 165-180. (First published 1974 in *Philosophical review* 83) Cambridge: Cambridge University Press.

Neher, Erwin and Bert Sakmann. (1992) 'The patch clamp technique', *Scientific american* 266: 28-35.

Niklasson, L. and N.E. Sharkey. (1992) 'Connectionism and the issues of compositionality and systematicity', in *Cybernetics and systems*, ed. R. Trappl. Dordrecht: Kluwer Academic Press.

Nolfi, Stefano and Domenico Parisi. (1992) 'Growing neural networks', Presented June 1992 at *Artificial life III* in Santa Fe, New Mexico.

O'Malley, Glenn. (1964) *Shelley and synesthesia*. Evanston, IL: Northwestern University Press.

Pandya, Deepak and Edward H. Yeterian. (1985) 'Architecture and connections of cortical association areas', in *Association and auditory cortices*, ed. Edward G. Jones and Alan Peters. New York: Plenum Press.

Paz, J.P.; S. Habib, and Wojciech H. Zurek. (1993) 'Reduction of the wave packet: preferred observable and decoherence time scale', *Physical review* D47: 488-501.

Paz, J.P. and S. Sinha. (1992) 'Decoherence and back reaction in quantum cosmology—Multidimensional minisuperspace examples', *Physical review* D45: 2823-2842.

Paz, J.P. and Wojciech H. Zurek. (1992) 'Environment induced superselection and the consistent histories approach to decoherence', *Los Alamos report no. LA-UR-92-878* (unpublished).

Penrose, R. (1985) 'Quantum gravity and state vector reduction', in *Quantum concepts in space and time*, ed. R. Penrosoe and C.J. Isham, pp. 129-146. Oxford: Oxford University Press.

Penrose, R. (1986) 'Big bangs, black holes, and "time's arrow"', in *Mindwaves*, ed. Raymond Flood and Michael Lockwood, pp. 259-276. Oxford: Basil Blackwell.
Penrose, R. (1989) *The emperor's new mind: Concerning computers, minds, and the laws of physics.* Oxford: Oxford University Press.
Pitcher, G., ed. (1968) *Wittgenstein: The philosophical investigations.* London: Macmillan.
Podolny, Roman. (1986) *Something called nothing.* Translated by Nicholas Weinstein. Moscow: Mir Publishers.
Pollack, Jordan B. (1992) 'Explaining cognition with nonlinear dynamics', Presented 9 September 1992 at the *First Swedish national conference on connectionism* in Skövde, Sweden.
Popper, K.R. and Eccles, J.C. (1977) *The self and its brain: An argument for interactionism.* Berlin: Springer.
Pour-El, Marian B. and J. Ian Richards. (1979) 'A computable ordinary differential equation which possesses no computable solution', *Annals of mathematical logic* 17: 61-90.
Pour-El, Marian B. and J. Ian Richards. (1981) 'The wave equation with computable initial data such that its unique solution is not computable', *Advances in mathematics* 39: 215-239.
Pour-El, Marian B. and J. Ian Richards. (1982) 'Noncomputability in models of physical phenomena', *International journal of theoretical physics* 21: 553-555.
Pour-El, Marian B. and J. Ian Richards. (1989) *Computability in analysis and physics.* Heidelberg: Springer-Verlag.
Puccetti, Roland. (1993) 'Dennett on the split-brain', *PSYCOLOQUY* 4(52) split-brain.1.puccetti.
Pugmire, D. (1989) 'Bat or batman', *Philosophy* 64: 207-217.
Purves, D. and J.W. Lichtman (1983) *Principles of neural development.* Sunderland, Massachusetts: Sinauer Associates.
Puthoff, Harold. (1989) *Physical review* A40: 4857+.
Puthoff, Harold. (1990) 'Everything for nothing', *New scientist* 127.
Putnam, Hilary. (1975) 'The meaning of "meaning"', *Minnesota studies in the philosophy of science* 7: 131-193.
Putnam, Hilary. (1988) *Representation and reality.* Cambridge, Massachusetts: MIT Press.
Pylyshyn, Zenon W. (1984) *Computation and cognition: Towards a foundation for cognitive science.* London: MIT Press.
Ralston, Zachary T. (1976) *Synesthesia in Gide's 'La symphonie pastorale'.* Citadel monograph series, no. 15. Charleston, South Carolina: Citadel, the Military College of South Carolina.
Ramsey, W.; S.P. Stich, and D. Rumelhart, eds. (1991) *Philosophy and connectionist theory.* Hillsdale, NJ: Lawrence Erlbaum Associates.
Rogers Jr., H. (1967) *Theory of recursive functions and effective computability.* New York: McGraw-Hill.
Rosch, E. and B.B. Lloyd (1978) *Cognition and categorization.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.
Rosenfield, I. (1988) *The invention of memory.* New York: Basic Books.
Rumelhart, David E. (1993a) Private communication 7 September at *Fifth JCI Summer School in Cognitive Science and Human Computer Interaction* at the University of Edinburgh, Scotland.
Rumelhart, David E. (1993b) Unpublished presentation September 1993 at *Fifth JCI Summer School in Cognitive Science and Human Computer Interaction* at the University of Edinburgh, Scotland.
Ryle, Gilbert. (1949) *The concept of mind.* London: Hutcheson.
Searle, J. (1992) *The rediscovery of mind.* Cambridge, Massachusetts: MIT Press.

Sharkey, N.E. (1992) 'The ghost of the hybrid: A study of uniquely connectionist representations', *Artificial intelligence and simulation of behaviour quarterly* 79: 10-16.
Shepherd, Gordon M., ed. (1990) *The synaptic organization of the brain,* 3 ed. Oxford: Oxford University Press.
Shoemaker, S. (1975) 'Functionalism and qualia', *Philosophical studies* 27: 291-315.
Shoemaker, S. (1981) 'Absent qualia are impossible: A reply to Block', *Philosophical review* 90: 581-599.
Skarda, C.A. and W.J. Freeman. (1987) 'How brains make chaos in order to make sense of the world', *Behavioral and brain sciences* 10: 161-195.
Smith, E.E. and D.L. Medin (1981) *Categories and concepts.* Cambridge, Massachusetts: Harvard University Press.
Smith, P. (1991) 'The butterfly effect', *Proceedings of the Aristotelian society* 91: 247-267.
Smith, P. (1993) *Chaos: Explanation, prediction & randomness.* Unpublished manuscript of Easter Term 1993 Cambridge Department of Philosophy lecture series.
Smith, P. and O.R. Jones. (1986) *The philosophy of mind.* Cambridge: Cambridge University Press.
Smolensky, P. (1990) 'Tensor product variable binding and the representation of symbolic structures in connectionist systems', *Artificial intelligence* 46: 159-216.
Sommerer, John C. and Edward Ott. (1993) 'A physical system with qualitatively uncertain dynamics', *Nature* 365: 138-140.
Sompolinsky, H. and A. Crisanti. (1988) 'Chaos in random neural networks', *Physical review letters* 61: 259-262.
Sperry, R.W. (1963) 'Chemoaffinity in the orderly growth of nerve fiber patterns and connections', *Proceedings of the national academy of sciences U.S.A.* 50: 703-710.
Sperry, R.W. (1965) 'Embryogenesis of behavioral nerve nets', in *Organogenesis,* ed. R.L. DeHaan and H. Ursprung, pp. 161-171. New York: Rinehart and Winston.
Stalnaker, R. (1993) 'Twin earth revisited', *Proceedings of the Aristotelian society* 63: 297-311.
Stone, Mark A. (1989) 'Chaos, prediction and LaPlacean determinism', *American philosophical quarterly* 26: 123-131.
Stühmer, W. (1991) 'Structure-function studies of voltage-gated ion channels', *Annual review of biophysics and biophysical chemistry* 20: 65-78.
Thomson, R.F. (1985) *The brain: An introduction to neuroscience.* W.H. Freeman & Company: New York.
Tsuda, Ichiro. (1994a) 'From micro-chaos to macro-chaos: Chaos can survive even in macroscopic states of neural activities', *PSYCOLOQUY* 5(12) eeg-chaos.3.tsuda.
Tsuda, Ichiro. (1994b) 'Can stochastic renewal of maps be a model for cerebral cortex?', *Physica D* (in press).
Tye, Michael. (1993) 'Blindsight, the absent qualia hypothesis, and the mystery of consciousness', in *Philosophy and the cognitive sciences,* ed. Christopher Hookway and Donald Peterson. Cambridge: Cambridge University Press. Note that this is the published version of 'Blindsight', presented September 1992 at the *Royal institute of philosophy: Philosophy and the cognitive sciences* conference in Birmingham, England.
Uttal, W.R. (1978) *The psychobiology of mind.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.
Uttal, W.R. (1981) *A taxonomy of visual processes.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.
van Essen, D.C. (1985) 'Functional organization of primate visual cortex', in *Cerebral cortex,* vol. 3, ed. A. Peters and E.G. Jones, pp. 259-329. New York: Plenum.
van Gulick, R. (1985) 'Physicalism and the subjectivity of the mental', *Philosophical topics* 13: 51-70.
von Neumann, J. (1955) *Mathematical foundations of quantum mechanics.* Princeton, NJ: Princeton University Press. (First German edition 1932).

von Neumann, J. (1956) 'Probabilistic logic and the synthesis of reliable organisms from unreliable components', in *Automaton studies*, ed. C. Shannon and J. McCarthy, pp. 43-98. Princeton: Princeton University Press.

Vrensen, G. and J. Nunes-Cardozo (1981) 'Changes in size and shape of synaptic connections after visual training: An ultrastructural approach to synaptic plasticity', *Brain research* 218: 79-97

Wagner, Sheldon; Ellen Winner, Dante Cichetti, and Howard Gardner. (1981) '"Metaphorical" mapping in human infants', *Child development* 52: 728-731.

Wheeler, Raymond Holder. (1920) *The synaesthesia of a blind subject.* Eugene, OR: The University Press.

Wheeler, Raymond Holder and Thomas D. Cutsforth. (1922) 'The synaesthesia of a blind subject with comparative data from an asynaesthetic blind subject', *University of Oregon publications* 1(10), June.

Wigner, E.P. (1962) 'Remarks on the mind-body question', in *The scientist speculates: An anthology of partly-baked ideas*, ed. I.J. Good, pp. 284-302. London: Heinemann.

Wigner, E.P. (1967) *Symmetries and reflections.* Bloomington, Indiana: Indiana University Press.

Wigstrom, H.; B.L. McNaughton, and C.A. Barnes (1982) 'Long-term synaptic enhancement in hippocampus is not regulated by post-synaptic membrane potential', *Brain research* 233: 195-199

Wilkins, Wendy K. and Jennie Wakefield. (forthcoming) 'Brain evolution and neurolinguistic preconditions', forthcoming in *Behavioral and brain sciences.*

Wilson, M.A. and J.M. Bower. (1989) In *Methods of neuronal modeling: From synapses to networks*, ed. C. Koch and I. Segev. Cambridge, Massachusetts: MIT Press.

Wilson, M.A. and J.M. Bower. (1992) 'Simulating cerebral cortical networks: Oscillations and temporal interactions in a computer simulation of piriform (olfactory) cortex', *Journal of neurophysiology* 67: 981-995.

Winograd, S. and J.D. Cowan (1963) *Reliable computation in the presence of noise.* Cambridge, Massachusetts: MIT Press.

Wittgenstein, L. (1953) *Philosophical investigations.* English 3d edition. New York: Macmillan.

Wolfram, Stephen. (1985) 'Origins of randomness in physical systems', *Physical review letters* 55: 449-452.

Wright, J.J. (1990) 'Reticular activation and the dynamics of neuronal networks', *Biological cybernetics* 62: 289-298.

Wright, J.J.; R.R. Kydd; and D.T.J. Liley. (1993) 'EEG models: chaotic and linear', *PSYCOLOQUY* 4(60) eeg-chaos.1.wright.

Wright, J.J.; R.R. Kydd; and D.T.J. Liley. (1994) 'Noise is crucial to EEG dynamics', *PSYCOLOQUY* 5 (19) eeg-chaos.5.wright.

Yao, Y. and W.J. Freeman. (1990) 'Model of biological pattern recognition with spatially chaotic dynamics', *Neural networks* 3: 153-170.

Zeki, S.M. (1975) 'The functional organization of projections from striate to peristriate visual cortex in the rhesus monkey', *Cold Springs Harbor symposia on quantitative biology* 40: 591-600.

Zeki, S.M. (1978) 'Functional specification of the cortex in the rhesus monkey', *Nature* 274: 423-428.

Zeki, S.M. (1981) 'The mapping of visual functions in the cerebral cortex', in *Brain mechanisms of sensation: Third Taniguchi symposium on brain sciences*, ed. Y. Katsuki, R. Norgren, and M. Sato, pp. 105-128. New York: Wiley.

Zipser, D. and D.E. Rumelhart. (1990) 'The neurobiological significance of the new learning models', in *Computational neuroscience*, ed. E.L. Schwartz, pp. 192-200. Cambridge, Massachussetts: MIT Press.

Zurek, Wojciech H. (1989) *Nature* 341: 119+.

Zurek, Wojciech H. (1991) 'Decoherence and the transition from quantum to classical', *Physics today* 44: 36-44.

Index